ATTITUDES AND SOCIAL COGNITION

INNOVATIONS IN SOCIAL PSYCHOLOGY

Targeted Identity-Safety Interventions Cause Lasting Reductions in Discipline Citations Among Negatively Stereotyped Boys

J. Parker Goyer and Geoffrey L. Cohen
Stanford University

Jonathan E. Cook
The Pennsylvania State University

Allison Master
The University of Washington

Nancy Apfel
Yale University

Wonhee Lee, Amelia G. Henderson,
and Stephanie L. Reeves
Stanford University

Jason A. Okonofua
University of California, Berkeley

Gregory M. Walton
Stanford University

High rates of discipline citations predict adverse life outcomes, a harm disproportionately borne by Black and Latino boys. We hypothesized that these citations arise in part from negative cycles of interaction between students and teachers, which unfold in contexts of social stereotypes. Can targeted interventions to facilitate identity safety—a sense of belonging, inclusion, and growth—for students help? Experiment 1 combined *social-belonging*, *values-affirmation*, and *growth-mindset interventions* delivered in several class sessions in 2 middle schools with a large Latino population ($N = 669$). This treatment reduced citations among negatively stereotyped boys in 7th and 8th grades by 57% as compared with a randomized control condition, 95% CI [-77%, -20%]. A growth-mindset only treatment was also effective (70% reduction, 95% CI [-84%, -43%]). Experiment 2 tested the social-belonging interven-

---

tion alone, a grade earlier, at a third school with a large Black population and more overall citations ($N = 137$ sixth-grade students). In 2 class sessions, students reflected on stories from previous 7th-grade students, which represented worries about belonging and relationships with teachers early in middle school as normal and as improving with time. This exercise reduced citations among Black boys through the end of high school by 65%, 95% CI [-85%, -15%], closing the disparity with White boys over 7 years by 75%. Suggesting improved interactions with teachers, longitudinal analyses found that the intervention prevented rises in citations involving subjective judgments (e.g., "insubordination") within 6th and 7th grades. It also forestalled the emergence of worries about being seen stereotypically by the end of 7th grade. Identity threat can give rise to cycles of interaction that are maladaptive for both teachers and students in school; targeted exercises can interrupt these cycles to improve disciplinary outcomes over years.

In 2013–2014, Black boys represented 7.9% of public school students but 25.2% of students suspended from school; Latino boys also receive elevated rates of discipline citations (U.S. Government Accountability Office, 2018; see also Losen, Hodson, Keith, Morrison, & Belway, 2015; Skiba et al., 2011; U.S. Department of Education, 2012). Discipline problems pose both an individual and a collective harm. They deny individual students opportunities to learn and contribute to inequality in adult life (Gregory, Skiba, & Noguera, 2010). They can have "collateral damage" on peers' achievement (Perry & Morris, 2014). And they undermine teachers' efforts, as the ability to create a classroom climate in which students stay on task is one of the strongest predictors of teachers' contribution to student learning (Ferguson, Phillips, Rowley, & Friedlander, 2015; Gates Foundation, 2010). Racial disparities in school discipline have persisted in cohort after cohort for many years (e.g., Losen et al., 2015). What causes high rates of discipline citations among Black and Latino boys as compared with White and Asian boys or girls generally?

At face value such large and persistent inequalities may seem to be primarily the consequence of stable features of the social context, such as violence or unpredictability in schools or in students' lives. They may also be attributed to stable characteristics of individuals, such as to bias in teachers, who may overestimate Black students' likelihood of misbehaving (Kunesh & Noltemeyer, 2015; Pigott & Cowen, 2000; see also Gershenson, Holt, & Papageorge, 2016; Quinn, 2017; Skiba et al., 2008), or to deficits in self-control in boys (e.g., Duckworth & Seligman, 2006) and the association between self-control and adolescent and adult outcomes (Mischel, Shoda, & Rodriguez, 1989; Moffitt et al., 2011).

We focus, however, on a dynamic social-relational process: how negative behavioral stereotypes influence students' and teachers' mutual perceptions and patterns of interaction in ways that undermine relationships over time (for a theoretical review, see Okonofua, Walton, & Eberhardt, 2016). Both students and teachers interact aware of stereotypes that allege that Black and Latino (hereafter *negatively stereotyped*) boys and men are violent, out-of-control, "troublemakers," or dangerous (Correll, Park, Judd, & Wittenbrink, 2002, Correll et al., 2007; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Gilliam, Maupin, Reyes, Accavitti, & Shic, 2016; Goff, Jackson, Di Leone, Culotta, & DiTomasso, 2014; Hugenberg & Bodenhausen, 2003; Kunesh & Noltemeyer, 2015; Okonofua & Eberhardt, 2015; Sagar & Schofield, 1980). Notably, although past research has tended to describe these stereotypes simply in racial terms, evidence is strongest for their application to Black boys and men, consistent with our focus on this race by gender subgroup. All but one of the studies cited above examined only perceptions of and responses to Black boys or men; the one that included female targets found disproportionate effects for perceptions of Black boys as compared with Black girls (Gilliam et al., 2016; see also Celious & Oyserman, 2001). Less research has examined stereotypes of Latinos as compared to Black boys, but some evidence points to similar stereotypes (Balderas, 2014; Gray, 2016; Gregory et al., 2010; Sadler, Correll, Park, & Judd, 2012; Skiba et al., 2011).

Past research shows that stereotypes can create self-fulfilling prophecies in social interaction, shaping perceptions of likability (Snyder, Tanke, & Berscheid, 1977), and performance on job interviews (Word, Zanna, & Cooper, 1974) and in academic testing situations (Jacoby-Senghor, Sinclair, & Shelton, 2016; Steele & Aronson, 1995). In real-world school settings, we theorize that stereotype-based concerns affect not just initial exchanges but the ongoing dynamics of interaction between students and teachers and ultimately disciplinary outcomes over years. Stereotypes may be especially relevant to school discipline, given the ambiguity in many teacher–student interactions. Stereotypes can shape how evaluators resolve ambiguity (Fiske & Taylor, 1991; Kunda, 1999), as when a teacher decides whether a bump was an act of aggression or playful horsing around (Duncan, 1976; Sagar & Schofield, 1980; see also Dovidio & Gaertner, 2000). Simultaneously, the perceived risk that social perception and judgments could be shaded by negative stereotypes can give rise to psychological threat and behaviors in students that become self-fulfilling (Steele, 1997), especially when this risk is ambiguous (Salvatore & Shelton, 2007).

Thus, we suggest that inequality in school discipline arises, in part, from small initial differences in social perception and interaction, which magnify over time as each party acts and reacts to the other. We test this question in two randomized field experiments. Can interventions that mitigate stereotype-based concerns and foster instead a sense of belonging, inclusion, and growth in students early in middle school forestall a negative cycle to improve the unfolding of social experience for students? Could this

approach partially remedy longstanding, high rates of disciplinary citations among Black and Latino boys over a meaningful period of time?

The transition to middle school is difficult for adolescents (Eccles et al., 1993). Instead of a personal relationship with one teacher and a single group of classmates as in elementary school, students change classes and thus know their teachers and classmates less well. Teachers also have higher academic expectations and punitive disciplinary policies become more commonplace, as teachers increasingly seek to maintain control. At the same time, students increasingly desire autonomy and respect, and may perceive teachers' disciplinary decisions as unfair (Yeager, Dahl, & Dweck, 2018). The structural changes of middle school, combined with the differing goals between students and teachers, create a context ripe for discipline problems.

Even as all students face these challenges, negatively stereotyped boys may experience the most significant difficulties (Simmons, Black, & Zhou, 1991), in part because the middle school context also enables stereotypes. Such stereotypes, we theorize, contribute both to identity threat among negatively stereotyped boys and to bias in teachers' disciplinary decisions (Okonofua, Walton et al., 2016).

As Black and Latino students enter adolescence they become increasingly aware of negative stereotypes about their group and reasonably apprehensive about the prospect of being seen and treated through the lens of these stereotypes (Cohen & Garcia, 2005; Yeager, Purdie-Vaughns, Hooper, & Cohen, 2017). Early negative experiences that appear to confirm these fears may prompt a wariness of teachers and reactive behavior. Notably, social threats, both in general and those that arise from a stigmatized group identity, can contribute to aggressive and antisocial behavior (e.g., Inzlicht & Kang, 2010; Twenge, Baumeister, Tice, & Stucke, 2001) and to social deviance (Belmi, Barragan, Neale, & Cohen, 2015; see also Tyler, 1990).

Simultaneously, the increasingly demanding and impersonal nature of middle school may raise the risk that teachers will in fact perceive and treat students in stereotypical ways. With more students to oversee, teachers have less time to focus on each individual student, reducing opportunities to individuate students, which can mitigate stereotyping (Locksley, Hepburn, & Ortiz, 1982; Pratto & Bargh, 1991; Rubinstein, Jussim, & Stevens, 2018). Teachers must also make many micro decisions rapidly, often under considerable pressure. Even when teachers do not endorse stereotypes explicitly (Warikoo, Sinclair, Fei, & Jacoby-Senghor, 2016; see also Quinn, 2017), stereotypes can function to create hypotheses or expectations (Darley & Gross, 1983). Pejorative expectations may facilitate the perception of misbehavior among Black and Latino boys and decisions to discipline them more frequently and more severely (Skiba et al., 2011; Skiba, Michael, Nardo, & Peterson, 2002). Such bias is especially likely as teachers respond to ambiguous incidents (Duncan, 1976; Sagar & Schofield, 1980; Skiba et al., 2002), or to repeated misbehaviors, which may be seen more rapidly as a problematic pattern when the student belongs to a negatively stereotyped group. Just a second minor infraction is enough for teachers to feel more troubled by a Black boy, to be more likely to identify him as a "troublemaker," and to desire to discipline him more severely than the same behavior from a White boy (Okonofua & Eberhardt, 2015).

Thus, stereotypes may fuel a self-perpetuating cycle of vigilance, perceptions of disrespect, and mistrust in middle school between Black and Latino boys, the primary target of discipline-related stereotypes, and teachers, the primary arbiters of disciplinary decisions (Okonofua, Walton et al., 2016; see also Gregory & Weinstein, 2008). If unchecked, this cycle may persist through high school and contribute to disparities in high school dropout (Hirschfield, 2009), postsecondary enrollment (Yeager, Purdie-Vaughns et al., 2017), and unemployment (Ritter & Taylor, 2011). This cycle becomes difficult to escape, as Black men are disproportionately more likely to be incarcerated than White men (Mc Kinnon & Bennett, 2005) and to return to prison after release (Jung, Spjeldnes, & Yamatani, 2010).

Within school settings, this cycle endures because of the perceptions and actions of both students and teachers. However, our theoretical approach should not be misunderstood as placing blame or responsibility (Okonofua, Walton et al., 2016). Whatever their causal role, in our theory neither children nor teachers are necessarily to blame for high levels of discipline citations; rather, a cycle of interactions fueled by a mutual awareness of negative stereotypes is at fault. As for partners in a distressed marriage (Gottman et al., 1976), toxic interactions become difficult to escape as patterns become entrenched. Yet this analysis implies that discipline citations may be reduced through early intervention with either teachers or students.

Focusing on teachers as an entry point to this system, one previous study encouraged middle-school math teachers to adopt an empathetic mindset toward misbehaving students—to understand and value students' perspective, even when this perspective seems unreasonable; to sustain positive relationships with students even when they misbehave; and to help students improve from within the context of such relationships (Okonofua, Paunesku, & Walton, 2016; see also Gregory et al., 2016). Implemented at five middle schools on the West Coast with a large Latino population, this intervention facilitated more respectful exchanges and halved suspension rates among racially diverse middle-school students over a school year as compared with students whose math teachers had received randomized control materials. Even as math teachers were the only teachers to be randomized, students of treated math teachers received lower suspensions from all of their teachers.

Another study examined an important exchange between students and teachers, the provision by teachers of critical academic feedback, at a school on the East Coast with a large Black population. Focusing on students who were performing moderately well in the fall of seventh-grade but could reasonably improve ("B" and "C" students, an 88-person subset of the full 277 correlational sample), researchers appended a note written by teachers to teachers' feedback on these students' essays about a personal hero in the spring semester of that year. The note conveyed that the teacher gave the feedback because he or she had high standards for the student and believed the student could reach those standards. This "wise feedback" note protected moderately well-performing Black students' trust in teachers in general over the rest of the seventh-grade school year (Yeager et al., 2014) and reduced discipline citations in eighth grade (Yeager, Purdie-Vaughns et al., 2017).

In both studies, a positive experience with a single teacher at a key developmental point improved students' interactions with

other teachers, reducing discipline problems that arose from interactions with, respectively, all teachers in the school during the same school year or grade-level teachers the next school year. These results imply that students' developing high-level beliefs about their social standing in school and relationships with teachers in general—beliefs relevant to their experience of identity threat—may contribute to discipline outcomes (see Gregory & Weinstein, 2008; Okonofua, Walton et al., 2016). Focusing on students, the present research tests whether promoting identity-safety directly in students could mitigate the negative cycles of interaction that give rise to sustained, high levels of discipline citations. We predicted that the benefits of this student-based approach would be greatest among Black and Latino boys, who experience the most severe behavioral stereotypes.[1] We thus implemented targeted exercises to promote beliefs in students that they can belong and learn in school and that their identity and values are accepted and welcome—what we call *identity safety*, drawing on past scholars (Purdie-Vaughns & Walton, 2011; Steele, 2010; Steele & Cohn-Vargas, 2013). If such exercises help negatively stereotyped boys reframe early interactions in middle school, they may improve students' early patterns of behavior, such as by reducing reactivity to negative or ambiguous exchanges. This improved behavior may foster better impressions in teachers, disconfirming negative stereotypes teachers may have, at least as applied to individual students. Teachers might then engage more positively with these students and give them the benefit of the doubt in future instances, forestalling the perception of a pattern of misbehavior (Okonofua & Eberhardt, 2015). As cycles of interaction improve, students may feel more confident in their belonging and less stereotyped (see Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009; Walton & Cohen, 2011), reinforcing a positive cycle that gives rise to lasting reductions in disciplinary citations.

At a high level, an important contribution of the present research is in testing the contribution of students' experience of social-identity threat to inequality in school discipline. We do so by taking a field experimental approach, testing whether targeted exercises to mitigate identity threat can change the system of interactions that produces high-levels of discipline problems in school. This approach complements teacher-focused approaches (Gregory et al., 2016; Hamre et al., 2012; Okonofua, Paunesku et al., 2016), and structural changes to support identity safety, such as ethnic studies curricula, which can also benefit Black and Latino adolescents (Dee & Penner, 2017), as well as school climate interventions (Bradshaw, Koth, Thornton, & Leaf, 2009) and relevant legislation (Rafa, 2018). Relative to approaches like these, the present research tests the specific causal role of students' experience of identity threat in the propagation of negative dynamics forward in time, and the potential for targeted exercises that address this experience to alter this system and improve an outcome of importance for both students and teachers.

By tracking outcomes up to seven years posttreatment, the present research also contributes to a question of broad importance for the field: How can brief psychological exercises, in at least some cases, cause significant, lasting change in individuals' lives, such as to raise achievement or improve college-going years into the future (Cohen et al., 2009; Goyer et al., 2017; Walton & Cohen, 2011)? Consistent with developing theory (e.g., Cohen & Sherman, 2014; Kenthirarajah & Walton, 2015; Walton & Wilson,

2018; Yeager & Walton, 2011), we propose in the context of school discipline that a shift in how students make sense of their experience in school need not stay isolated in the minds of students. Instead, a change in meaning and interpretation can alter students' developing social experience. By altering students' behavior in the presence of and toward teachers, and teachers' responses to that behavior, identity-safety interventions may improve ongoing (or "recursive") cycles in a critical relationship. This approach draws on research on close relationships, which finds that interventions to alter how people make sense of marital conflicts (Finkel, Slotter, Luchies, Walton, & Gross, 2013) or their partner's feelings for them (Marigold, Holmes, & Ross, 2007, 2010) can instigate cycles of behavior that improve relationships over time (see also Gottman et al., 1976). We suggest that change in relationships can serve as an important vehicle for lasting change in areas of people's lives beyond the relationship (e.g., Newcomb, Koenig, Flacks, & Warwick, 1967). Consistent with our theorizing, teacher expectancies are thought to cause lasting gains in student learning by altering cycles of interaction between teachers and students, where higher expectations, better teaching, and greater care from teachers produce greater engagement and learning in students (Rosenthal & Jacobson, 1968). Likewise, improvement in students' approach to learning as induced by a psychological intervention can be noticed by teachers (Blackwell, Trzesniewski, & Dweck, 2007), positioning them to respond positively, sustaining change (Skinner & Belmont, 1993). Moreover, in examining interactions with new teachers year upon year, we suggest that identity-safety interventions can improve patterns of interactions with important types of relationship partners, not just with a specific relationship partner (see Yeager, Purdie-Vaughns et al., 2017).

This approach also extends past research examining the mechanisms that can lead identity-safety interventions in particular, such as social-belonging and values-affirmation interventions, to cause lasting gains. Past research has focused on how these interventions can mitigate cycles of threat and poor performance (Cohen et al., 2009), forestall global interpretations of adversities (Walton & Cohen, 2007, 2011; Sherman et al., 2013; Yeager, Purdie-Vaughns et al., 2017), promote engagement with peers and

---

[1] Our focus on Black and Latino boys contrasts with past research but this difference reflects specific methodological decisions. Okonofua and colleagues (2016) found that the teacher treatment lowered suspension rates across student Race × Gender groups, even as the observed reduction was from a higher base for Black and Latino boys, consistent with national data (U.S. Government Accountability Office, 2018). All students may benefit from a teacher who responds more positively in discipline situations. Although Yeager and colleagues' (2017) wise-feedback note might also be expected to produce the largest reduction in disciplinary citations for Black boys, a reduction for all Black students was found. However, by including only students who had earned a B or C GPA in the previous semester, the study disproportionately excluded Black boys with low grades and these students also had especially high baseline levels of disciplinary citations. Reanalysis of original data from the full sample ($N = 277$; osf.io/3hpu8) shows that, in sixth grade, Black boys had more discipline citations than any other Race × Gender group, consistent with the present research, yet those Black boys who participated in the wise-feedback study had fewer sixth-grade citations than those who did not participate ($M_{participants} = 2.41$; $M_{nonparticipants} = 3.65$). Had those Black boys been included in the experimental study, they may have shown especially large reductions if assigned to treatment. The present studies, by contrast, included participants regardless of baseline indices.

potential mentors (Walton & Cohen, 2007; Walton, Logel, Peach, Spencer, & Zanna, 2015; Yeager, Walton et al., 2016), and help students enter more advanced academic tracks (Goyer et al., 2017). We contribute a focus on recursive processes in teacher–student relationships.

We report two studies. Experiment 1 combines three identity-safety interventions, *social belonging* (Walton & Cohen, 2011), *growth mindset* (Paunesku et al., 2015), and *values affirmation* (Cohen et al., 2009), and tests this combined treatment on disciplinary citations in a relatively large sample over 2 years, during the seventh and eighth grades. It also tests the growth-mindset intervention as a standalone treatment. Experiment 2 tests the social-belonging intervention alone. Although it uses a smaller sample, Experiment 2 focuses on the population at greatest risk, Black boys (U.S. Government Accountability Office, 2018). It examines, moreover, a context in which students interact with an almost-all White teaching staff, as is the case in many schools (Gay & Howard, 2001) but one that may exacerbate racial tensions and miscommunications (Cohen, Steele, & Ross, 1999; Dee, 2004; Gershenson, Hart, Lindsay, & Papageorge, 2017; Pigott & Cowen, 2000; Wright, 2016). Experiment 2 also features longitudinal analyses over a longer (7-year) assessment period, sixth grade through 12th grade, to better understand effects over time. A contribution of this research is in testing whether existing interventions that, in previous research, have primarily been examined in terms of one outcome (achievement) can also affect other important outcomes (discipline; see Gregory et al., 2010). We also adapt the social-belonging intervention, which has previously been tested only with college students (e.g., Walton & Cohen, 2011), to examine its potential benefit for early adolescents.

Experiments 1 and 2 were conducted as separate longitudinal field experiments and thus had somewhat different goals and designs. Yet we report them together as they inform each other and their samples and results are complementary. Although Experiment 1 was conducted second and Experiment 2 first, we present them in reverse order for ease of exposition. In both studies we were most interested in effects on negatively stereotyped boys, given behavior-related stereotypes and their disproportionate rates of disciplinary citations, but we also explored effects for other groups.

Both studies provide a window into the recursive social-relational and psychological processes that can allow a change initially triggered by a brief, time-limited intervention to last. First, we tested whether the interventions would reduce subjective citations early in middle school, which otherwise may instigate a negative recursive cycle thereafter. Relative to objective citations (e.g., "being tardy"), subjective citations (e.g., "insubordination") require teachers to exercise judgment regarding both whether an infraction occurred and, if so, its severity. Past research shows that racial inequalities are substantially larger for subjective than for objective citations (Skiba et al., 2002; see also Yeager, Purdie-Vaughns et al., 2017). Subjective citations may both reflect mistrust that has built up between students and teachers and deepen that mistrust, and as such contribute to the recursion of discipline problems over time. If the interventions mitigate this cycle, early effects may emerge on subjective citations.

Second, both studies include specific analyses to further inform how the interventions may mitigate negative discipline cycles. These analyses differ across studies so as to be appropriate to each context (overall level of discipline citations and distribution of subjective vs. objective citations) and design (assessment period, sample size). In Experiment 1, where the sample is larger but rates of citations are lower, we tested whether the intervention reduced the risk that one subjective citation led to another in seventh grade, the year the intervention was delivered, and if it reduced the risk that students who received one or more subjective citations in seventh grade had any citations (subjective or objective) the next school year (cf. Okonofua & Eberhardt, 2015). In Experiment 2, the unprecedented temporal window, a 7-year assessment period, higher rates of citations, and a more even distribution of subjective and objective citations (in Experiment 1, most citations were subjective) allowed a more nuanced examination of patterns within and across school years. If negative dynamics emerge over time as Black boys and specific (mostly White) teachers interact, these may manifest as rising numbers of subjective citations within the school year early in middle school. An intervention that remedies a toxic cycle may thus have the greatest effect in reducing subjective citations later in the school year. In addition, Experiment 2 examined corresponding psychological changes in students, including whether the intervention reduced uncertainty about belonging in students (i.e., reduced the degree to which students reported that challenges in school made them feel they do not belong, Walton & Cohen, 2007), forestalled worries about being seen and treated in light of negative stereotypes (cf. Cohen et al., 2009; Walton & Cohen, 2011), and promoted students' belief in their overall level of belonging in their school (Walton & Cohen, 2011; Walton et al., 2015).

Finally, an important question involves the role of the social context. Experiences of identity threat arise from contexts; mitigating threat then allows students to pursue whatever opportunities are available to them in the context they experience. Identity-safety interventions are not magic bullets. They will not always produce the same gains; instead, consistent with the Lewinian tradition in social psychology, they reveal the constraints and affordances of a social system (see Cohen & Sherman, 2014; Walton & Wilson, 2018; Yeager & Walton, 2011). In an initial effort to explore the role of the social context, the two schools featured in Experiment 1 and the school featured in Experiment 2 differed along a variety of dimensions, including region of the country, socioeconomic status of students, primary race-ethnicity of students, and overall level of disciplinary citations. In the General Discussion section we return to the role of the social context and discuss potentially important contextual moderators of treatment effects (see also Borman, Grigg, Rozek, Hanselman, & Dewey, 2018; Broda et al., 2018; Hanselman, Rozek, Grigg, & Borman, 2017; Protzko & Aronson, 2016; Walton et al., 2015).

Experiment 1

Experiment 1 tested whether social-belonging, growth-mindset, and values-affirmation interventions could, in combination, reduce discipline problems for negatively stereotyped boys. Although these interventions developed separately in the literature, they address interrelated processes, each of which contributes to students' overall experience of psychological threat or safety in school (for comparisons of these interventions, see Garcia & Cohen, 2012; Rattan, Savani, Chugh, & Dweck, 2015; Walton, Paunesku, & Dweck, 2012; Yeager & Walton, 2011). Each helps

students represent themselves in school in ways broader than a psychological threat would allege—as more than a token of a group who might not belong or as a person of fixed and low academic ability. Instead, they invite students to see themselves as adequate and capable, as people with enduring values who can connect, grow, and belong. In this sense, each encourages an "expandable self," countering the narrow representation of the self implied by a psychological threat (Walton et al., 2012). Each, moreover, can facilitate students' sense of belonging in school over time (see Walton & Brady, 2017). Yet they do so in somewhat different ways.

The *social-belonging* intervention features survey results and stories from older students to convey that challenges and worries about belonging are normal in an academic transition and improve with time, not evidence of a permanent lack of belonging. Students then reflect on their own experience and how it reflects this process of change, often in a message or letter for future students. The exercise gives students a more adaptive narrative with which to make sense of adversities they face. Past studies of students in the transition to college find that the treatment helps students sustain a sense of belonging in the face of day-to-day adversities, stay academically engaged in college, and build stronger peer and mentor relationships (Walton & Cohen, 2011; Walton et al., 2015; Yeager, Walton et al., 2016). A contribution of the present research is in adapting the intervention for middle school and testing whether the more hopeful lens it provides for making sense of early challenges in middle school can set in motion a cycle of better interactions with teachers to cause lasting reductions in discipline citations.

The *growth-mindset* intervention conveys that intelligence is not fixed but can grow with hard work, good strategies, and help from others, often featuring findings from neuroscience about the malleability of the brain and the metaphor that the brain can grow "like a muscle" (e.g., Aronson, Fried, & Good, 2002; Blackwell et al., 2007; Paunesku et al., 2015). Rather than negative group-laced inferences that can arise from social adversities ("I/people like me don't belong here"), growth-mindset interventions aim to remedy negative inferences that can emerge from academic setbacks (e.g., "I'm bad at school") by representing academic improvement as possible. Although research on growth and fixed mindsets did not develop specifically to understand the effects of negative stereotypes, worries about ability and about belonging can feed on one another in environments defined by achievement (e.g., school) to form a toxic cycle and contribute to identity threat. Indeed, the perception that peers and instructors in school hold a fixed theory of intelligence can, in combination with perceptions of negative stereotypes about the intelligence of specific groups, contribute to feelings of nonbelonging among people subject to those stereotypes (Good, Rattan, & Dweck, 2012; see also Emerson & Murphy, 2015; Rattan et al., 2018). Moreover, growth-mindset exercises, especially when administered in classroom settings, may seem to reflect the core values of the teacher or the institution and thus reassure students that stereotypes of fixed inability will not be applied to them there (see Emerson & Murphy, 2015; Yeager, Walton et al., 2016). Although growth mindset interventions have not focused on discipline outcomes, other research finds that representations of growth and the possibility of improvement can reduce discipline problems for Black boys (Oyserman, Terry, & Bybee, 2002).

Values-affirmation interventions, like social-belonging interventions, were developed specifically to understand and mitigate threat that arises from negative stereotypes (Cohen, Garcia, Apfel, & Master, 2006). However, instead of providing students an adaptive narrative with which to make sense of adversities in school (e.g., "It is normal to worry about belonging at first"), values-affirmation interventions seek to bolster students' sense of personal adequacy to help them cope with identity threat (e.g., "I am ok" or "I am good enough"; for reviews, see Cohen & Sherman, 2014; Garcia & Cohen, 2012). They do so through exercises that invite students to reflect on and write about personally important values (Cohen et al., 2006, 2009). Reflecting on core values reminds students of unconditional sources of worth, restoring a sense of personal security and adequacy that helps them confront threats more effectively. Notably, this enhanced sense of security can forestall global interpretations of day-to-day adversities in school (Sherman et al., 2013) and support students' sense of belonging (Cook, Purdie-Vaughns, Garcia, & Cohen, 2012)—processes that parallel those set in motion by the social-belonging intervention (Walton & Cohen, 2007, 2011; Walton et al., 2015).

Experiment 1 tested an "omnibus" condition, combining all three interventions. It seemed possible that this combined treatment might be more effective than any treatment on its own, either by reaching a broader group of students or by increasing the impact on individual students. We also included a growth-mindset-only condition to begin to test the effects of each treatment, and in part given the larger goals of that particular study to understand achievement gaps more broadly. (Experiment 2 tests a social-belonging only condition.)

The intervention was delivered in six class sessions over the first two thirds of the seventh-grade school year. In the combined condition, there were two doses per component (social belonging, growth mindset, and values affirmation), with one dose (a single 15- to 20-min session) of each delivered in each semester of seventh grade. Seventh grade—rather than sixth grade, as in Experiment 2—was chosen because (a) this is when past growth-mindset (Blackwell et al., 2007) and values-affirmation interventions (Cohen et al., 2009) have been delivered effectively in middle school; (b) seventh grade is associated with rising social and academic challenges and growing racial disparities in school trust and concerns about being seen stereotypically (Yeager, Purdie-Vaughns et al., 2017), a pattern also seen in Experiment 2 (see Figure 6); and (c) pilot testing raised concerns about students' ability to self-administer reading-and-writing exercises in sixth grade in this population.

The two middle schools that participated in Experiment 1 differed in important ways from the school that participated in Experiment 2 (see Table S1). Mostly prominently, they were located in the Western rather than the Northeastern United States and, as provided by publicly available data, were lower in socioeconomic status (more students received free or reduced lunch), contained a primarily Latino rather than Black racial-ethnic student sample, more of the teachers were non-White, and fewer had a Master's degree or higher. There were also fewer disciplinary citations overall. Among control-condition students, the Experiment 1 schools had, on average, 0.68 citations per year per student in seventh and eighth grade versus 2.38 in Experiment 2 (see Table S25). A similar proportional difference was present among control-condition negatively stereotyped boys (M_{Exp1} = 1.50 vs.

$M_{\text{Exp2}} = 6.45$). Further, the citations at Experiment 1's schools were mostly subjective in nature (77% of citations at one school site and 78% at the other) rather than evenly divided between subjective and objective as in the Experiment-2 school (where 53% of all citations issued were subjective).

Method

Participants and school settings. The two middle schools that participated in Experiment 1 met specific screening criteria. These criteria ensured that the schools included large numbers of low-income and racial-minority students and were large and stable enough to support a research study. Each had similar total enrollment, a sizable population of negatively stereotyped students (as noted, mostly Latino), a proportion of students receiving free and reduced lunch greater than 30%, less than 10% student mobility rate, and clear racial achievement gaps in academic performance. At the time of recruitment, both schools had an Academic Performance Index lower than 800, the target score the state required to demonstrate adequate academic performance and progress.

The two schools in Experiment 1 had slightly different demographic profiles (see Table S1). Middle School 1 ($n = 341$ randomly assigned student participants) contained 64% negatively stereotyped students (students who were not White or Asian), and nearly half (48%) of the student body received free or reduced lunch. Middle School 2 ($n = 401$ randomly assigned student participants) contained 42% negatively stereotyped students. Fewer of its students received free and reduced lunch (34%), and it had a higher academic progress score (795 vs. 782 for Middle School 1 in the year the schools were recruited, 2 years prior to the start of the study).

We obtained consent from nearly the entire cohort in the relevant grade (seventh grade) in each school (Middle School 1: 96%; Middle School 2: 94%). Negatively stereotyped students in Experiment 1 were primarily Latino ($n = 305$; 41% of the total sample combining middle school sites) along with a smaller number of Black students ($n = 38$) and students of other non-Asian racial-ethnic groups ($n = 29$), including American Indian or Alaskan Native ($n = 14$), Native Hawaiian or other Pacific Islander ($n = 4$), and those who identified with two or more races ($n = 11$). Given our interest in mitigating discipline citations among all racial-ethnic students who face negative discipline stereotypes (even as these may be strongest for Black boys), analyses combined these groups into a single category. This negatively stereotyped group included 193 boys (160 Latino, 16 Black, and 17 from other non-Asian, non-White ethnicities).

The nonstereotyped category included White ($n = 264$) and Asian ($n = 106$) students across both genders. Asians were grouped with Whites because they are not subject to the same negative intellectual or behavioral stereotypes as other racial-ethnic groups (e.g., see Quinn, 2017). Table S3 shows the race-ethnicity of students by condition, for the full sample and for each middle school separately.

Procedure. All participants and their parent/guardian(s) provided informed consent to participate in the study. Parent consent forms were either mailed or sent home with students prior to the new school year. In an effort to remove barriers to participation for low-socioeconomic status students, the procedure relied upon passive consent; only students whose parents returned the signed opt-out form were excluded from the study. Student assent was also obtained for each of the intervention activities. All procedures and data collection were approved by the Stanford University Institutional Review Board and by institutional agreements with the partner schools.

At the start of seventh grade, students were randomly assigned to one of three conditions, with approximately one third of students assigned to each condition. Each condition involved six class sessions, three per semester: (a) a control condition ($n = 246$) in which students were provided neutral activities during all six sessions; (b) a combined condition ($n = 248$) in which students were provided exercises for values affirmation, growth mindset, and social belonging, in two sessions, six sessions in total; and (c) a growth-mindset-only condition ($n = 248$) in which students were provided the growth-mindset intervention in two sessions and neutral activities in the other four sessions.

Random assignment was conducted using a stratified random sampling technique: Students were randomized to condition within Race × Gender groups within seventh-grade classrooms, within school. Reflecting school preferences, students were randomized in science classrooms at Middle School 1 and in social studies classrooms at Middle School 2. Teachers were blind to student condition.

One "dose" of each intervention was delivered in both the fall semester (Grading Periods 1 and 2) and spring semester (Grading Periods 3 to 5, those after the winter holiday break) of seventh grade, in a cyclical fashion. The order each semester was values affirmation, growth mindset, and social belonging. Interventions were introduced cyclically, rather than consecutively, after pilot research suggested the former would increase student engagement. Values affirmation was introduced first because prior studies show it is most effective early in the school year, prior to any decline in performance (Cook et al., 2012). First-semester doses were administered in Weeks 3, 9, and 13 of the school year; this encompassed the first 30% of the school year. Second-semester doses were administered in Weeks 22, 25, and 29; this was from 51% to 67% of the school year. Second-semester doses were compressed to maximize the potential impact on discipline outcomes over the remainder of the school year. See Table S4 for a detailed study timeline.

Retention of participants. Of the 742 students originally randomly assigned, 722 students had nonmissing seventh-grade discipline data (712 students had nonmissing eighth-grade discipline data). Retention rates (i.e., number of students with nonmissing outcome data) did not differ by condition (see online supplementary materials). We defined the primary sample as participants with available outcome data who saw the assigned randomized materials given in all three sessions in Semester 1 of seventh grade (i.e., at least the first three of six exercises administered during the seventh-grade school year, $N = 669$). This was the minimum number of doses required for a combined-condition student to have seen materials for all three interventions that comprised that condition, and also ensured that the growth-mindset only condition saw materials for at least one of the two growth-mindset intervention sessions. Students who did not see this material were generally absent and unavailable for make-ups. This primary sample comprised 175 negatively stereotyped boys, 153 negatively stereotyped girls, 172 nonstereotyped boys, and 169 nonstereotyped girls.

Of students who saw materials for the first three sessions, 95% ($n = 637$; 86% of the total sample) completed essay prompts in all three sessions with appropriate responses (i.e., indicating they had read and comprehended the materials; inappropriate responses were blank or did not answer the question; see online supplementary materials). A total of 621 students (93% of the primary sample) saw assigned randomized materials in all six sessions offered across Semesters 1 and 2 in seventh grade. Of these, 94% ($N = 581$; 87% of the primary sample) completed all six sessions with appropriate responses. Though, as these descriptive statistics indicate, students generally provided appropriate responses, these were not used to determine selection into the sample, which was based only on taking part in the first three intervention sessions, as described above.

Identity safety interventions. Each intervention session was brief (15–25 min per session), administered in class, and involved a personalized, interactive reading-and-writing exercise. The social-belonging and growth-mindset interventions were delivered online; the values-affirmation exercise was distributed in class in sealed envelopes marked with each student's name and collected by teachers at the end of the session in resealed envelopes (using the same procedure as Cohen et al., 2009). While teachers administered the materials and oversaw the intervention sessions, they were blind to their content and to individual students' condition assignment.

Values affirmation comprised two 15-min writing exercises in which students reflected on their core values and wrote about why those values were important to them (Cohen et al., 2009). The growth-mindset intervention, adapted from Paunesku et al. (2015), comprised two 20-min sessions that used examples from neuroscience and other students, along with guided writing prompts, to convey that intelligence can grow with hard work, effective strategies, and help from others. The social-belonging intervention was similar to that used in Experiment 2 with modest tailoring for the two school contexts in Experiment 1 informed by focus groups with older students at these schools. Described in more detail in Experiment 2, for which the middle-school version was first developed, it comprised two 25-min sessions in which students read brief stories about how older students initially worried about fitting in at middle school and how these concerns dissipated with time. See online supplemental materials for further description and Tables S5 to S7 for sample student responses to essay prompts within each intervention exercise.

Each control session involved exercises matched in length and interactivity to the treatment exercises; these were adapted from control materials used in past intervention trials. The values-affirmation control condition involved reflecting on how values not important to the self might be important to others (Cohen et al., 2009). The growth-mindset control condition involved learning about how scientists study the brain and the functions of its four lobes (Paunesku et al., 2015). The social-belonging control condition conveyed that most students initially have little interest in state government when they enter seventh grade but come to care more about it as they spend more time in middle school (developed in Experiment 2).

Dependent measures. We tracked discipline citations over 2 years, during the seventh-grade intervention year and through eighth grade (i.e., the remainder of middle school).

Average annual total citations in seventh and eighth grade. The primary outcome was the average annual discipline citations a student received in seventh and eighth grade (average total citations per year over two years), as provided by official records from the school district that oversaw both schools. We represented this outcome as an annual average rather than a sum because a few students left the school district each year (see Tables S8 and S9). We also examined the total number of citations students received in each grade separately.

Average annual subjective citations in seventh and eighth grade. We coded the brief description the school district provided for each citation to categorize citations as subjective or objective, drawing from classifications used in past research (Skiba et al., 2002) and policies specified in the discipline handbooks from this and other school districts (we intended our coding scheme to be inclusive enough to permit its application beyond the schools in the present sample). We used the same coding scheme across both experiments (see Tables S10–S11).

Subjective citations (e.g., "insubordination" or "disrespect"; see Gregory & Weinstein, 2008; Skiba et al., 2011, 2002) involved both interactive and ambiguous behaviors that required discretion on the part of the person observing or targeted by the behavior (a teacher or other school official) in determining whether a violation of a discipline policy had occurred. For these events, observers could disagree about whether an infraction had occurred and, if so, the severity of the consequence it warranted (they thus had to make a "judgment call"; see Yeager, Purdie-Vaughns, et al., 2017). Additionally, subjective citations often involve routine interpersonal interactions in the classroom and could thus be influenced by a teacher's preexisting relationship with a student and, in turn, influence that relationship going forward. Thus, from the student perspective, subjective citations might seem to be associated with an inconsistent application of rules by teachers.

Objective citations (e.g., "being tardy" or "cutting class") typically reflected the violation of a specific or objective rule that was also associated with a specific consequence, requiring less discretion and thus increasing consistency in disciplinary judgments. These citations might also be observed by a wider range of staff members than classroom teachers. At the Experiment 1 schools, nearly 80% of the total citations issued at each school were subjective. Although this distribution prohibits comparisons of subjective versus objective citations in this sample, it means that overall analyses involve primarily subjective citations. To facilitate comparison with Experiment 2, we also examine treatment effects on subjective citations alone.

Baseline measures.

Baseline disciplinary behavior. In both experiments, we used disciplinary outcomes in the school year immediately prior to the intervention as a covariate in analyses to increase the precision of treatment estimates (due to random assignment, this covariate was balanced between conditions, as explained further below). As Experiment 1's intervention was administered in seventh grade, the relevant baseline covariate was total sixth-grade citations (Experiment 2 used fifth-grade citations), $M = 0.37$, $SD = 1.68$, $Min = 0$, $Max = 31$. Because the measure was positively skewed (skewness = 11.59; Kurtosis = 184.06), we tried various transformations appropriate for positive skewness to reduce the skew. As in Experiment 2, a log transformation (natural log of $1 +$ the original variable) resulted in the greatest reduction in positive

skew (skewness = 3.35; Kurtosis = 16.65). The resulting variable had a mean of 0.16 and standard deviation of 0.42, with the maximum reduced to 3.47. (Experiment 2 also includes teacher ratings of preintervention classroom behavior as a covariate; these were not available in Experiment 1.)

Though analyses controlled only for baseline disciplinary behavior, we also used the following variables to assess baseline equivalence.

Baseline academic performance. Baseline academic performance was students' core course GPA in both semesters of sixth grade, $M = 3.15$, $SD = 0.95$, $Min = 0.33$, $Max = 4.33$. Data were available for 95% ($N = 638$) of the three-session primary sample ($N = 669$).

Baseline socioeconomic status. Two measures of socioeconomic status were available, whether students received free or reduced lunch (38% did), and students' reports of their parents' level of education, measured on a 1 (*not a high school graduate*) to 5 (*graduate school*) scale, $M = 3.38$, $SD = 1.30$, $Min = 1$, $Max = 5$. Free and reduced lunch status was available for all students in the primary sample; parental education was available for 88% of the primary sample ($N = 589$).

Baseline equivalence. To assess the success of random assignment, we regressed each baseline covariate and demographic measure on two dummy codes, one for each treatment comparison (= 1) with the control condition (= 0), for the full sample and for each Race $\times$ Gender subsample. We used logistic regression for the dichotomous variables representing race, gender, and free or reduced lunch status, linear regression for parental education and sixth-grade GPA, and negative binomial regression for sixth-grade discipline citations. In the full sample (i.e., the sample including all four Race $\times$ Gender groups), there were no differences between the control condition and either of the two treatment conditions on any baseline covariate, zs < 1.50, ps $\geq .15$. See Table S12. For negatively stereotyped boys, the focal group of interest, there were no significant differences between the control and either treatment condition for any baseline covariate (all ps $\geq .22$, except for sixth-grade GPA, which was marginally higher in the combined condition relative to the control condition, $b = 0.33$, $t(162) = 1.79$, $p = .075$). See Table S13.

We additionally tested for equivalency of random assignment by regressing our principal outcome, annual average discipline citations during seventh and eighth grade, on the four baseline covariates simultaneously (free and reduced lunch status, parental education, sixth-grade GPA, and sixth-grade discipline citations). We then regressed the predicted values from this regression model on each treatment versus control dummy code. This technique is used to reduce the possibility of detecting significant differences in baseline characteristics by chance, relative to when testing multiple covariates separately, by examining whether condition groups differ in the portion of the outcome variance that is most correlated with the baseline covariates. When we did this, there was no main effect of condition for either treatment, for either the full sample or negatively stereotyped boys specifically, ts < 1, ps ≥ 0.34.

Data analyses. The article text reports results for the three-session intent-to-treat sample ($N = 669$), which we call the primary sample. The supplement provides results for the original sample ($N = 742$) and six-session intent-to-treat sample ($N = 621$). Results are generally consistent across sample definitions.

We used negative binomial regression to test the treatment effect. This approach is appropriate and conservative, as it accounts for the non-normal distribution of discipline citations and in so doing restricts the influence of extreme values (see Atkins, Baldwin, Zheng, Gallop, & Neighbors, 2013; Atkins & Gallop, 2007). It is more appropriate than Poisson regression, another technique used to analyze effects on count variables, when the dependent count variable is overdispersed (i.e., the variance of the outcome is greater than the mean, indicating there is more variability in the outcome than a Poisson distribution implies). In negative binomial regression, outcome values are hypothesized to emerge from a Poisson process, but the inclusion of an overdispersion parameter, alpha, permits variation in the outcome to be greater than in a true Poisson process. Across analyses, likelihood ratio tests comparing the Poisson and negative binomial models favored the negative binomial model (consistently, the null hypothesis that the alpha parameter was zero, as in a Poisson model, was rejected). Using a negative binomial approach also has the advantage of maintaining consistency with the analytical approach in Experiment 2, where the overdispersion in discipline outcomes is even greater. We represent the effect size of coefficients in negative binomial models as an incident rate ratio (IRR), as is standard for this kind of regression. The IRR is a proportional change that is the exponentiated form of the corresponding log-count regression coefficient (alternatively, the IRR $-$ 1 expresses the same ratio as a percentage change). For treatment effects for a given Race $\times$ Gender group, the IRR is the proportional change in (adjusted) citations in the treatment condition relative to the control condition.

We predicted treatment effects for negatively stereotyped boys. Given that interactions between race, gender, and condition can only be tested in a full sample analysis, we use full-sample analyses to report high-level summative results. However, because of the nature of the disciplinary outcome, which in both experiments was high for negatively stereotyped boys and low for other groups, it is also reasonable to test the treatment effect for this group in subsample analyses, which makes the fewest assumptions about the equivalence of groups. Subgroup analyses are appropriate when one group is identified prospectively as having a much higher risk than others of a negative outcome (here, a higher mean level of discipline citations) in the absence of treatment (Moyé & Deswal, 2001; Rothwell, 2005). When an unequal distribution of risk is present, a full sample analysis, which assumes that variance is equal across groups, can result in inaccurate standard errors. This is a more significant issue in Experiment 2, where the disparity in citations between negatively stereotyped boys and other groups is greater: In the seventh and eighth grades, raw average annual citations for control-condition negatively stereotyped boys were 6.45 (vs. an average of 1.43 for the other three Race $\times$ Gender groups), a difference of 5.02 citations; in Experiment 1, the same disparity was 1.10 ($M_{\text{NegStBoys}} = 1.50$ vs. $M_{\text{Others}} = 0.40$). However, for completeness, we also report the critical contrasts for negatively stereotyped boys on high-level outcomes in Experiment 1 obtained from subsample analyses. For the same reason, the longitudinal analyses in the article text examining change in citations within and/or across school years (i.e., Experiment 2), which focus primarily on treatment effects for Black boys, feature results from subsample analyses (the online supplemental materials provides detailed results for all Race $\times$

Gender groups). In Experiment 1, these analyses examine recursive processes among negatively stereotyped boys.

The main effect of condition was represented as two dummy codes, in the same model: combined treatment ($= 1$) versus control ($= 0$), and growth-mindset-only treatment ($= 1$) versus control ($= 0$). We also adjusted for total citations in the school year prior to intervention, centered on the mean for each Race $\times$ Gender group, so that adjusted means would characterize the average student for each group (see Goyer et al., 2017; Sherman et al., 2013). To assess treatment effects on high-level outcomes within the full Race $\times$ Gender sample, we initially included main effects of race, gender, and their two- and three-way interactions with each condition contrast (as in Experiment 2 below). We ultimately excluded the three-way interactions from the final full-sample model, as this provided a better-fitting model for Experiment 1's data (which, as noted, featured a smaller disparity between negatively stereotyped boys and other groups).

Race ($0 =$ negatively stereotyped, $1 =$ nonstereotyped), gender ($0 =$ male, $1 =$ female), and each treatment versus control condition variable ($0 =$ control, $1 =$ treatment) were dummy-coded so the coefficient associated with a given condition variable would represent the relevant treatment effect for negatively stereotyped boys, and the intercept (expressed as an incident rate ratio) would represent the adjusted mean citations for negatively stereotyped boys in the control condition. The treatment effect (i.e., simple effect of condition) for each of the other three Race $\times$ Gender groups was obtained by defining dichotomous 0/1 race and gender variables for the given group as 0, using the same regression model. In addition to comparing each treatment with the control condition, we also compared each treatment to the other. All analyses tested effects when collapsing across the two schools by including a dichotomous indicator for school site, which was standardized so the intercept coefficient would represent the combined-school average for the reference category (i.e., the Race $\times$ Gender $\times$ Condition group defined as 0 on relevant variables).

First we report the effect of both intervention conditions for each Race $\times$ Gender group on the primary summative outcome, average annual total citations during seventh and eighth grade. Then we report results in each grade separately. We also report within-grade treatment effects on subjective citations specifically. Next we use logistic regression analyses among Black boys to begin to understand whether the treatments mitigated negative recursive cycles, by first reducing additional subjective citations in seventh grade and then by eliminating citations in eighth grade among those negatively stereotyped boys who received at least one subjective citation in seventh grade. For all negative binomial analyses, we report the log-count coefficient b, the exponentiated log-count coefficient (the IRR) along with its 95% confidence limits (which in turn are the log-count 95% confidence limits exponentiated), and also express the IRR as a percentage reduction in citations relative to the reference category in the given regression model (IRR $- 1$).

Results

Summative analyses: Average annual citations in seventh and eighth grade. Planned contrasts obtained from the full-sample analysis revealed the predicted reduction in discipline citations for negatively stereotyped boys. The combined intervention reduced average annual total citations over the last two years of middle school for this group by 57%, $b = -0.84$, $z = -2.69$, $p = .01$, IRR $= 0.43$, 95% CI [0.23, 0.80], $M_{\text{adj-C}} = 1.20$ versus $M_{\text{adj-COMB}} = 0.52$ (means represent the adjusted number of citations for each condition). The growth-mindset-only condition also produced a significant (70%) reduction, $b = -1.21$, $z = -3.70$, $p < .001$, IRR $= 0.30$, 95% CI [0.16, 0.57], $M_{\text{adj-C}} = 1.20$ versus $M_{\text{adj-GM}} = 0.36$. The two treatments did not differ (i.e., the combined condition vs. growth-mindset-only dummy code was nonsignificant), $b = 0.38$, $z = 1.11$, $p = .27$, IRR $= 1.46$, 95% CI [0.75, 2.83]. See Figure 1.

In the subsample analysis, the reduction in average annual total citations for negatively stereotyped boys was 52% in the combined condition, $b = -0.74$, $z = -2.31$, $p = .02$, IRR $= 0.48$, 95% CI [0.26, 0.90], $M_{\text{adj-C}} = 1.18$ versus $M_{\text{adj-COMB}} = 0.56$, and 64% in the growth-mindset-only condition, $b = -1.03$, $z = -3.03$, $p = .002$, IRR $= 0.36$, 95% CI [0.18, 0.70], $M_{\text{adj-C}} = 1.18$ versus $M_{\text{adj-GM}} = 0.42$. As in the full sample analysis, the two treatments did not differ, $b = 0.29$, $z = 0.83$, $p = .41$, IRR $= 1.34$, 95% CI [0.67, 2.68].

Although our primary interest was in treatment differences among negatively stereotyped boys, we also examined higher order interactions in the full-sample model. There was a significant Gender $\times$ Condition interaction for both the combined condition contrast, $b = 0.96$, $z = 2.08$, $p = .038$, IRR $= 2.60$, and the growth-mindset-only condition contrast, $b = 0.98$, $z = 2.06$, $p = .04$, IRR $= 2.66$, suggesting reductions occurred for boys and not girls. However, the reductions for *nonstereotyped* boys were only directional, 29% for the combined condition and 21% for the growth-mindset-only condition (see online supplemental materials). Neither comparison reached significance ($|z|s \leq 0.80$, $ps \geq .42$), likely in part because the number of citations for nonstereotyped boys in the control condition was already low, only 0.25 citations. Girls of both racial-ethnic groups who, like nonstereotyped boys, had low levels of citations in the control condition, showed no change as a result of either treatment, $|z|s < 1.45$, $ps \geq .16$. The Race $\times$ Condition interaction was only significant for the growth-mindset-only condition contrast, $b = 0.98$, $z = 2.04$, $p = .041$, IRR $= 2.66$. Incident rate ratios for all regression coefficients in the full-sample model are shown in the middle panel of Table S14.

Total citations within grade.

Seventh-grade citations. Figure 1A displays adjusted citations in seventh grade, the year of intervention delivery, for each Race $\times$ Gender group as a function of condition. For negatively stereotyped boys, total seventh-grade citations were reduced by 55% in the combined condition, $b = -0.79$, $z = -1.80$, $p = .072$, IRR $= 0.45$, 95% CI [0.19, 1.07], $M_{\text{adj-C}} = 1.36$ versus $M_{\text{adj-COMB}} = 0.62$, and by 75% in the growth-mindset only condition, $b = -1.39$, $z = -3.11$, $p = .002$, IRR $= 0.25$, 95% CI [0.10, 0.60], $M_{\text{adj-C}} = 1.36$ versus $M_{\text{adj-GM}} = 0.34$. There were no differences between the two treatments in seventh-grade citations, $b = 0.60$, $z = 1.30$, $p = .19$, IRR $= 1.83$, 95% CI [0.74, 4.56].

In the subsample analysis (i.e., only among negatively stereotyped boys), the combined-condition reduction in total seventh-grade citations was 58% and significant, $b = -0.87$, $z = -1.99$, $p = .046$, IRR $= 0.42$, 95% CI [0.18, 0.99], $M_{\text{adj-C}} = 1.36$ versus $M_{\text{adj-COMB}} = 0.57$. The growth-mindset-only condition reduction

A

B
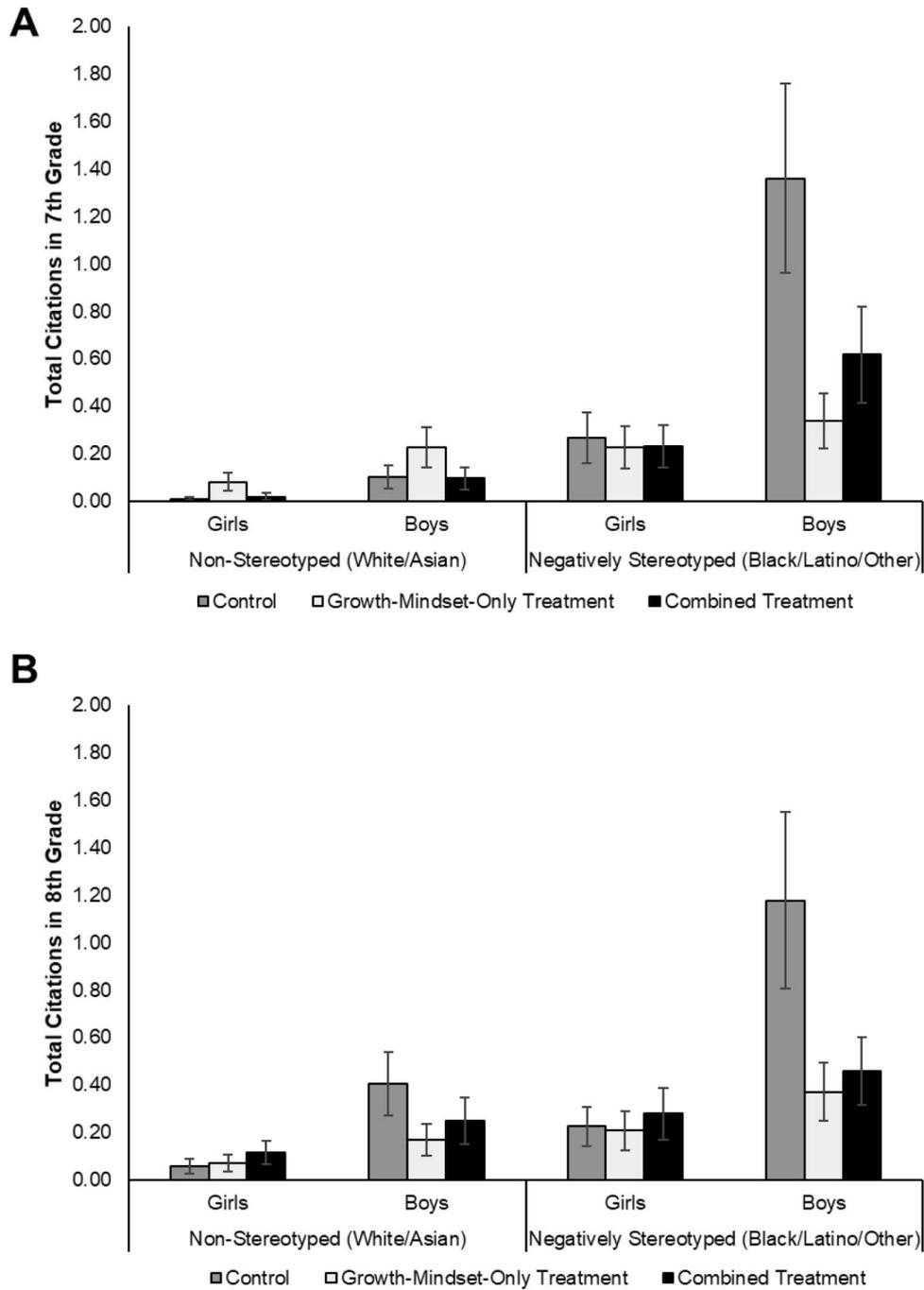
Figure 1. The adjusted average number of annual discipline citations during (A) seventh and (B) eighth grade for nonstereotyped (White/Asian) and negatively-stereotyped (Black/Latino/Other) students in Experiment 1, as a function of gender and condition. The sample includes all students with available outcome data who saw materials for the three exercises administered in Semester 1 of seventh grade ($N = 669$ for seventh-grade citations; $N = 663$ for eighth-grade citations). Means were obtained from single-level full-sample negative binomial regression analyses that collapsed across the two middle-school sites and adjusted for preintervention (sixth grade) discipline citations. Error bars represent ± 1 SE. Combined treatment = social belonging + values affirmation + growth mindset. For the regression model from which these means are derived, see the middle panel of Table S14.

was 70% and also significant, $b = -1.21$, $z = -2.58$, $p = .01$, IRR = 0.30, 95% CI [0.12, 0.75], $M_{adj-C} = 1.36$ versus $M_{adj-GM} = 0.40$. As in the full sample analysis, the two treatments did not differ, $b = 0.34$, $z = 0.72$, $p = .47$, IRR = 1.40, 95% CI [0.56, 3.53].

Examining the other Race × Gender groups using the same full-sample model, there were no significant differences in citations between the control condition and either treatment in seventh grade for negatively stereotyped girls, $|z|s < 0.35$, $ps > .75$, or for nonstereotyped boys, $|z|s < 1.40$, $ps \geq .17$. Although there was a significant increase in citations for nonstereotyped girls in the growth-mindset-only condition relative to the control condition, $b = 2.03$, $z = 2.69$, $p = .007$, IRR = 7.59, 95% CI [1.74, 33.20], this result, which was not anticipated, reflects the extremely low base in the control condition; citations were still near zero in both conditions, $M_{adj-C} = 0.01$ versus $M_{adj-GM} = 0.08$.

Eighth-grade citations. Figure 1B displays adjusted citations in eighth grade, the first full postintervention year, for each Race × Gender group as a function of condition, again obtained from a full-sample model. Full-sample results were consistent with those in seventh grade, and in some cases, stronger. For negatively stereotyped boys, total eighth-grade citations were reduced by 61% in the combined condition, $b = -0.95$, $z = -2.17$, $p = .030$, IRR = 0.39, 95% CI [0.17, 0.91], $M_{adj-C} = 1.18$ versus $M_{adj-COMB} = 0.46$, and by 68% in the growth-mindset only condition, $b = -1.16$, $z = -2.58$, $p = .010$, IRR = 0.32, 95% CI [0.13, 0.76], $M_{adj-C} = 1.18$ versus $M_{adj-GM} = 0.37$. Again, the two treatments did not differ significantly, $b = 0.21$, $z = 0.48$, $p = .63$, IRR = 1.23, 95% CI [0.52, 2.91].

In the subsample analysis, the combined-condition reduction in total eighth-grade citations for negatively stereotyped boys was 45% and nonsignificant, $b = -0.60$, $z = -1.34$, $p = .179$, IRR = 0.55, 95% CI [0.23, 1.32], $M_{adj-C} = 0.99$ versus $M_{adj-COMB} = 0.55$. The growth-mindset-condition reduction was 57% and marginally significant, $b = -0.84$, $z = -1.85$, $p = .064$, IRR = 0.43, 95% CI [0.18, 1.05], $M_{adj-C} = 0.99$ versus $M_{adj-GM} = 0.43$. As in the full sample analysis, the two treatments did not differ, $b = 0.25$, $z = 0.54$, $p = .59$, $IRR = 1.28$, 95% CI [0.52, 3.15].

For nonstereotyped boys, the growth-mindset only condition resulted in a marginal (58%) reduction in eighth-grade citations, $b = -0.87$, $z = -1.76$, $p = .079$, IRR = 0.42, 95% CI [0.16, 1.11], $M_{adj-C} = 0.41$ versus $M_{adj-GM} = 0.17$. The combined condition produced a similar but smaller and nonsignificant reduction (39%), $b = -0.49$, $z = -0.99$, $p = .32$, IRR = 0.61, 95% CI [0.24, 1.60], $M_{adj-C} = 0.41$ versus $M_{adj-COMB} = 0.25$. There were no significant differences between the control condition and either treatment in eighth grade for girls of either racial-ethnic group, $zs < 1.15$, $ps \geq .26$.

Subjective citations within grade. To examine subjective citations, we used the same full-sample model as above, but excluded objective citations from the outcome. In general, treatment effects were stronger when examining subjective citations alone.

Seventh-grade subjective citations. For negatively stereotyped boys, full-sample analyses indicated that subjective seventh-grade citations were reduced by 71% in the combined condition, $b = -1.25$, $z = -2.69$, $p = .007$, IRR = 0.29, 95% CI [0.12, 0.71], $M_{adj-C} = 1.13$ versus $M_{adj-COMB} = 0.32$, and by 85% in the growth-mindset-only condition, $b = -1.88$, $z = -3.89$, $p < .001$,

$IRR = 0.15$, 95% CI [0.06, 0.39], $M_{adj-C} = 1.13$ versus $M_{adj-GM} = 0.17$. These treatment effects did not differ significantly from each other, $b = 0.64$, $z = 1.24$, $p = .21$, IRR = 1.89, 95% CI [0.69, 5.14].

In the subsample analysis, the combined-condition reduction in subjective seventh-grade citations for negatively stereotyped boys was also 71% and significant, $b = -1.23$, $z = -2.74$, $p = .006$, IRR = 0.29, 95% CI [0.12, 0.70], $M_{adj-C} = 1.09$ versus $M_{adj-COMB} = 0.32$. The growth-mindset-only condition reduction was 80% and significant, $b = -1.60$, $z = -3.31$, $p = .001$, IRR = 0.20, 95% CI [0.08, 0.52], $M_{adj-C} = 1.09$ versus $M_{adj-GM} = 0.22$. As in the full sample analysis, the two treatments did not differ, $b = 0.37$, $z = 0.72$, $p = .47$, IRR = 1.44, 95% CI [0.53, 3.89].

There were no treatment effects on subjective seventh-grade citations for negatively stereotyped girls or nonstereotyped boys, $zs < 0.90$, $ps \geq .38$. As with total citations, for nonstereotyped girls there was an unexpected significant difference in subjective citations between the growth-mindset-only and control condition, $b = 2.31$, $z = 2.67$, $p = .008$, but adjusted means for all condition groups were near zero ($\leq.05$ subjective citations).

Eighth-grade subjective citations. For negatively stereotyped boys, subjective eighth-grade citations were reduced by 63% in the combined condition, $b = -0.99$, $z = -2.35$, $p = .019$, IRR = 0.37, 95% CI [0.16, 0.85], $M_{adj-C} = 0.83$ versus $M_{adj-COMB} = 0.31$, and by 72% in the growth-mindset-only condition, $b = -1.29$, $z = -2.91$, $p = .004$, IRR = 0.28, 95% CI [0.12, 0.66], $M_{adj-C} = 0.83$ versus $M_{adj-GM} = 0.23$. Again, the two treatments did not differ significantly, $b = 0.30$, $z = 0.67$, $p = .50$, IRR = 1.36, 95% CI [0.56, 3.29].

In the subsample analysis, the combined-condition reduction in subjective eighth-grade citations for negatively stereotyped boys was 52% and nonsignificant, $b = -0.73$, $z = -1.64$, $p = .100$, IRR = 0.48, 95% CI [0.20, 1.15], $M_{adj-C} = 0.75$ versus $M_{adj-COMB} = 0.36$. The growth-mindset-condition reduction was 65% and significant, $b = -1.06$, $z = -2.28$, $p = .022$, IRR = 0.35, 95% CI [0.14, 0.86], $M_{adj-C} = 0.75$ versus $M_{adj-GM} = 0.26$. Despite this, the two treatments still did not significantly differ from each other, $b = 0.34$, $z = 0.69$, $p = .49$, $IRR = 1.40$, 95% CI [0.54, 3.61].

For nonstereotyped boys, subjective citations were reduced by 63% in the growth-mindset-only condition, $b = -1.00$, $z = -1.98$, $p = .048$, IRR = 0.37, 95% CI [0.14, 0.99], $M_{adj-C} = 0.29$ versus $M_{adj-MIND} = 0.11$. Nonstereotyped boys exhibited a nonsignificant (52%) reduction in the combined condition, $b = -0.73$, $z = -1.44$, $p = .15$, IRR = 0.48, 95% CI [0.18, 1.31], $M_{adj-C} = 0.29$ versus $M_{adj-COMB} = 0.14$. There were no condition differences in subjective eighth-grade citations for girls of either racial-ethnic group, $zs < 0.75$, $ps \geq .46$.

Recursive processes. Leveraging the relatively large sample and appropriate for the relatively low overall rates of citations in Experiment 1, we used logistic regression to examine whether the interventions mitigated negative recursive cycles where an initial subjective citation leads to more citations. These analyses focused on the subsample of negatively stereotyped boys who received a first subjective citation in seventh grade (23% of negatively stereotyped boys, $N = 40$).

Specifically, we sought to determine, first, if either intervention reduced the odds that these students received one or more additional subjective citation that school year, preventing an

exacerbating cycle. Second, we tested whether either intervention reduced the odds that they received any citations in eighth grade, cutting off a negative discipline cycle entirely. We report adjusted means as probabilities and represent effect sizes as odds ratios (OR). In seventh grade, the effect size was the odds of receiving two or more subjective citations versus one, in the relevant treatment condition (combined or growth-mindset-only) relative to the control condition. In eighth grade, it was the odds of receiving any (subjective or objective) citations versus none, in the relevant treatment condition versus the control condition. In each analysis, we controlled for sixth-grade citations and school site, as in other subsample analyses for the primary high-level outcomes.

Odds of receiving an additional subjective citation in seventh grade. Were negatively stereotyped boys who received a first subjective citation in seventh grade less likely to receive another, if they were assigned to an identity-safety intervention? Yes, for those students assigned to the combined intervention. The combined treatment reduced the odds that negatively stereotyped boys received one or more additional subjective citations by 85%, a 50% reduction in probabilities from 82% to 41%, $b = -1.91$, $z = -2.03$, $p = .042$, $OR = 0.15$, 95% CI [0.02, 0.94], $M_{\text{adj-C}} = 0.82$ versus $M_{\text{adj-COMB}} = 0.41$. The growth-mindset-only treatment also reduced the odds of receiving an additional subjective citation (by 80%), a 42% reduction in probabilities, but this did not reach significance, $b = -1.63$, $z = -1.54$, $p = .123$, $OR = 0.20$, 95% CI [0.02, 1.56], $M_{\text{adj-C}} = 0.82$ versus $M_{\text{adj-GM}} = 0.48$. See Figure 2.

Odds of receiving at least one citation in eighth grade. While the effect of the treatment in mitigating a recursive process in seventh grade was most evident for subjective citations, the effect into eighth grade extended to all citations. The combined treatment

Figure 2. The figure depicts adjusted probabilities from a subsample logistic regression analysis involving negatively stereotyped boys who received at least one subjective citation in seventh grade, as a function of condition, controlling for total sixth-grade citations and school site. Those in the combined treatment had significantly lower odds than otherwise identical controls of receiving an additional subjective citation in seventh grade, and of receiving any citation in eighth grade. The corresponding reductions in the growth-mindset-only condition did not reach significance. For more detail on the regression model from which these means were derived, see Table S19.

significantly reduced the odds that negatively stereotyped boys who had received at least one subjective citation in seventh grade received any citations (subjective or objective) in eighth grade by 91%, a 69% reduction in probabilities, $b = -2.44$, $z = -2.44$, $p = .014$, $OR = 0.09$, 95% CI [0.01, 0.62], $M_{\text{adj-C}} = 0.78$ versus $M_{\text{adj-COMB}} = 0.24$. In the growth-mindset-only condition, the reduction in odds was slightly smaller (73%; equivalent to a 36% reduction in probabilities) and did not reach significance, $b = -1.29$, $z = -1.28$, $p = .199$, $OR = 0.27$, 95% CI [0.04, 1.98], $M_{\text{C}} = 0.78$ versus $M_{\text{MIND}} = 0.50$. See Figure 2 and Table S19.

Discussion

Using a relatively large sample, Experiment 1 found that targeted in-class exercises to mitigate students' experience of identity threat in middle school reduced discipline citations among negatively stereotyped boys over two years. The treatment combined social-belonging, values-affirmation, and growth-mindset interventions delivered across six class sessions in seventh grade. Negatively stereotyped boys provided this combined treatment as compared with control materials for at least the first three of these sessions showed a 57% reduction in annual discipline citations in seventh and eighth grades. However, Experiment 1 also suggests that all three interventions in combination may not be necessary to achieve meaningful reductions in discipline problems; indeed, a single targeted intervention may be sufficient, as a growth-mindset-only intervention also reduced discipline citations (by 70%) among negatively stereotyped boys. Across all primary analyses, this treatment could not be statistically distinguished from the combined treatment. As most citations in Experiment 1 were subjective in nature, both treatment effects primarily involved reductions in incidents requiring the subjective judgment of teachers.

Logistic-regression analyses focused on those children most at risk suggest that the combined treatment in particular cut off a negative recursive cycle among negatively stereotyped boys in interaction with their teachers: Among negatively stereotyped boys who, in seventh grade, already received one subjective citation, the combined treatment reduced the probability of receiving an additional subjective citation by 50%. For fewer students did one subjective citation lead to another. The combined treatment also reduced the probability, by 69%, that these same students had any citations in eighth grade. The same effects were smaller in magnitude and nonsignificant for the growth-mindset-only condition (42% and 36% respectively). It is possible that this intervention achieved disciplinary reductions through a somewhat different mechanism than social belonging and values affirmation interventions, which target identity threat directly. For example, the growth-mindset-only intervention did reduce the total number of citations at-risk negatively stereotyped boys received in eighth grade (see Table S20), suggesting that it may have prevented a negative cycle from becoming worse, even as it did not cut off the cycle entirely. It also reduced the number of eighth-grade citations received by nonstereotyped boys, who presumably do not experience threat due to negative stereotypes.

Experiment 2

Experiment 2 tests the effect of the social-belonging intervention alone in mitigating discipline citations, focusing on the group
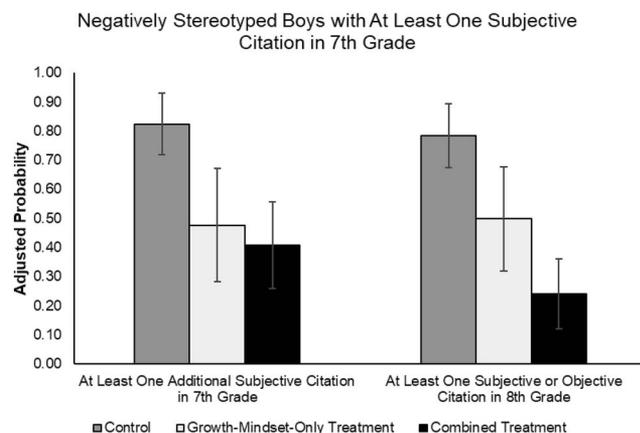
at greatest risk for high levels of citations, Black boys. We examined this intervention alone because it most directly addresses students' beliefs about their developing relationships with teachers, which we hypothesize to be critical to the emergence of discipline patterns over time. Because the social belonging intervention focuses specifically on easing academic transitions, we delivered it as students made the transition to middle school, at the beginning of sixth grade.

Experiment 2 also provides an extended longitudinal assessment: a 7-year period instead of two. The 7-year period allowed us to assess whether the social belonging intervention causes reductions in discipline citations into and throughout high school. Such a pattern would imply the persistent operation of students' beliefs and altered patterns of interaction with teachers. If, instead, effects taper off at the transition to high school, it would imply that the effects were tied to the specific middle-school context, and students' experiences and established patterns of interaction there.

Because disciplinary citations in Experiment 2 are both higher overall and more evenly divided between subjective and objective citations (53% vs. 44%; 3% were unspecified), Experiment 2 also allows for a more nuanced examination of treatment effects on each type of citation over time, both within and across grade levels in middle school. Specifically, we assess: Does the social belonging intervention prevent within-year increases in subjective citations? This analysis builds on but goes beyond the recursive analysis in Experiment 1, by examining rates of citations, not the odds of an additional citation (due to the greater rate of citations in Experiment 2); by examining patterns over six semesters (fall and spring of the sixth, seventh, and eighth grades) rather than annually across 2 years; and by examining whether the effect is specific to subjective citations, by direct comparison with objective citations, using the same regression model (due to their greater presence in Experiment 2).

Furthermore, unlike Experiment 1, Experiment 2 examines students' psychological experience in middle school both absent intervention and when given the narrative that early adversities and worries about belonging are normal in middle school and improve with time. Experiment 2 sought to assess whether this narrative helps Black boys feel less uncertain of their belonging—that is, whether it reduced the extent to which they questioned their belonging in the face of challenges. Over time, the intervention should forestall students' worry about being seen and treated in light of a negative stereotype and bolster students' belief about their overall level of belonging, both because the intervention should improve cycles of interaction with teachers and should prevent students from drawing negative global inferences from any specific negative incidents that do occur (see Yeager, Purdie-Vaughns et al., 2017).

Finally, as noted, Experiment 2 also examines a rather different school context as compared to Experiment 1: A smaller school on the East Coast rather than the West Coast, with a predominant Black rather than Latino racial-ethnic group of interest, and higher overall rates of disciplinary citations (see Table S1 and Table S25).

Method

Participants and setting. The study began in a middle school in a middle-class community in the Northeastern United States. Almost all teachers (93%) at the school were White, as in many schools (Gay & Howard, 2001). The primary sample consisted of all Black and White sixth-grade students at the school who provided consent to participate in the study and who saw randomized material ($N = 137$), approximately half of the school's sixth-grade class. This lower consent-rate for Experiment 2 relative to Experiment 1 is likely due to a different consent process. Experiment 2 used an opt-in approach (see the Procedure section below).

However, data from earlier cohorts at the same school prior to intervention suggest that, at least along important indices of academic achievement, students participating in research through this consent process do not differ from nonparticipating students. Prior research finds that the poor performance rate for Black students (percentage receiving a D or below) prior to intervention was similar to the randomized control conditions in prior years when the school participated in studies involving values affirmation (see supplement, Cohen et al., 2006).

The sample was evenly divided among four Race × Gender groups: Black boys ($n = 25$), Black girls ($n = 43$), White boys ($n = 39$), and White girls ($n = 30$). Though we do not have descriptive data on students whose parents did not consent, the school-wide student population at the time of the study was about half Black, as in our sample. This sample yielded 12 Black boys in the control condition and 13 in the treatment condition.

Beyond the 137 Black and White students included in the primary sample, a smaller number of Asian (one girl) and Latino (four boys, seven girls) students also took part (see Table S31). Because these students were present in such low numbers (constituting only 8% of the student body), their data are not included in primary analyses, allowing us to focus on Black and White children (see Simmons et al., 1991). However, because Latino boys can also face negative stereotypes and heightened discipline citations (e.g., Experiment 1), we examined the robustness of the high-level intervention effect (i.e., summative effect, averaging across years) in secondary analyses with the slightly larger sample ($N = 149$) that combines Black and Latino boys ($n = 29$), Black and Latino girls ($n = 50$), White and Asian boys ($n = 39$), and White and Asian girls ($n = 31$).

Statistical power and approach. The sample size in Experiment 2 was limited, constrained primarily by the use of only one school site, and consent obtained from only about half of all enrolled students in the grade of interest (sixth grade). We took several steps to increase power. First, we used repeated measures, which can increase power through both the frequency and the duration of measurements (Gibbons et al., 1993; Open Science Collaboration, 2015; Raudenbush & Xiao-Feng, 2001). Our primary outcome, annual discipline citations, was assessed in seven measurements over 7 years. Subjective and objective citations in middle school were assessed six times over 3 years (every semester), and social-psychological measures were assessed five times over 3 years. Second, controlling for predictive covariates, such as fifth-grade disciplinary citations and sixth-grade preintervention psychological beliefs, increased power by reducing unexplained error variance and increasing the precision of treatment estimates (e.g., Light, Singer, & Willett, 1990; Murnane & Willett, 2011). Third, power can be increased by using reliable outcome measures. Our primary outcome is a behavioral measure that reflects all times each student was recorded as having committed a discipline violation. As such, it is less likely than many measures used in psychological research to have high levels of measurement error.

Fourth, we randomized students to condition within Race × Gender blocks, removing the portion of the residual variation that is between blocks. This increase in precision is greatest when the variation between blocks is expected to be large, as here (discipline citations are greater for negatively stereotyped boys than for other groups under the status quo, as described previously). Finally, power is higher for larger effect sizes. Our intervention was a high-impact, interactive exercise, versions of which were previously shown to cause large improvements in school experience and academic outcomes that persist over years (e.g., Walton & Cohen, 2011; Walton et al., 2015). Moreover, we customized this intervention for our sample through focus groups and surveys conducted with older students at the same school, and had it delivered by teachers in classroom sessions, increasing its experimental realism and impact (see Aronson & Carlsmith, 1968). Considering these issues, as well as the difficulty of obtaining such a diverse longitudinal field sample, we felt justified in pursuing our research questions with this dataset.

Even as we used these strategies to increase power, we also used statistical techniques that ensure that extreme values do not drive the effects. As in Experiment 1, analyses were conducted using negative binomial regression, an appropriate and conservative approach to these data relative to techniques like linear or Poisson regression, as described below.

Procedure. All participants and their parent/guardian(s) provided informed consent to participate in the study. In contrast with Experiment 1, active consent (using an opt-in procedure) was required: Permission slips were distributed to students' parents in the mail or through their students' classroom. The study only included students who returned signed permission slips indicating that their parents allowed them to participate in the study. All procedures and data collection were approved by the Yale University Institutional Review Board and by institutional agreements with the partner schools. Consenting students were randomly assigned to either the control condition or the social-belonging intervention, approximately half to each condition, within Race × Gender blocks, within classroom. The intervention was administered as an in-class exercise in two 25-min sessions in the first quarter of sixth grade, in late September and a month later. Both consenting and nonconsenting students completed the activity as a class exercise; nonconsenters completed control materials and their data were never collected. While teachers administered the materials, they were blind to their content and to individual students' condition assignment, as in Experiment 1. Teachers distributed sealed envelopes marked with each student's name and collected completed materials in resealed envelopes.[2]

Discipline and other administrative data were provided by the school district from third through 12th grade. Thus, students' discipline outcomes were tracked for 7 years, for as long as they attended the middle school site (Grades 6–8) and the district's mainstream high school (Grades 9–12). In addition, we administered a standard social-psychological "climate assessment" (paper-and-pencil survey) at the beginning and the end of the sixth, seventh, and eighth grades with the first assessment preceding the treatment.

Retention of participants. Because the study was conducted in a single relatively stable school district (i.e., few students tended to leave the district from year to year) and the primary outcome, discipline citations, could be obtained from the school site rather than from students themselves (e.g., as is necessary for self-report measures), retention from semester to semester was generally high, as shown in Table S32. A total of 71% of the full Black and White student sample ($N = 137$) was retained through 12th grade. Of those who left ($N = 40$), 43% left during middle school (during Grades 6 to 8) and 58% left during high school (during Grades 9 to 12). Leaving the district did not differ by condition, $zs < 0.50$, $ps > .65$, for the entire sample or for Black boys specifically, $zs < 0.70$, $ps \geq .50$. Retention rates were similar for the larger sample of Black, Latino, White, and Asian students (see Table S32).

We had some discipline data over the 7-year assessment period for every student in the sample, enabling us to compute an annual average for all 137 Black and White students in the primary sample. Nearly all students (99%, or $N = 135$) had some middle school discipline data. One Black boy and one Black girl, both in the control condition, did not have middle school data. Both returned to the district in high school and thus contributed to all analyses that used high school data. A large proportion (81%, or $N = 111$) of students in the primary sample also had some high school discipline data (see middle panel of Table S30). This included 92% of Black boys in the control condition and 85% of Black boys in the treatment condition. The presence versus absence of disciplinary records in high school did not differ by condition overall or for Black boys specifically, $zs < 0.85$, $ps > .40$. In addition (see bottom panel of Table S30), 128 students (93% of the primary sample) had data for one or more psychological "climate assessments" administered at the start and end of each school year in middle school (i.e., a baseline at the start of sixth grade, outcomes from the end of sixth grade to the end of eighth grade). Of those missing psychological data (only 7% of the full sample), four were in the control condition and five were in the treatment condition.

Students missing data on some postintervention measures were included wherever possible; namely, except for analyses that required data at specific measurement occasions where they were missing. This approach is simple and transparent and retains our experimental design. Students with discipline data from some but not all years in a given assessment period were retained in single-level analyses of average annual discipline citations in that period. They also were included in all longitudinal analyses, contributing to the regression estimates in periods when they had data. Similarly, for social-psychological outcomes, students with any data contributed to relevant longitudinal analyses in the time periods when they had data.

Social-belonging intervention. The intervention featured stories and conclusions drawn from interviews and surveys conducted with seventh-grade students at the same school in the previous school year about their transition to this school and were presented

---

[2] At the start of seventh-grade, students were rerandomized as part of another study testing the effects of values affirmation. This exercise had no main effect on disciplinary outcomes, collapsing across belonging condition, for any subgroup. As will be seen, by seventh grade the belonging intervention had already initiated change in discipline outcomes and relevant psychological processes for Black boys. Whether social-belonging and values-affirmation interventions are redundant in this context or can be additive is an important question for future research (see Experiment 1; Walton et al., 2015). Unfortunately, the sample size in Experiment 2 was not large enough to support interaction tests. However, examining the direction of means, there was no pattern of note.

as such (e.g., The intervention stated "Last year the seventh-grade students at [school name] answered questions about how they felt about [school name] when they were in the sixth-grade like you"). Using a version of the intervention that addressed worries about belonging in the transition to college as a model (Walton & Cohen, 2011), the materials conveyed that it is normal to worry about belonging and relationships with teachers upon entering middle school but that these concerns lessen with time. Even as they addressed common social difficulties, the materials did not deny the possibility of bias or that students' experiences may differ along racial or gender lines. As in past versions, the exercise included a brief "saying-is-believing" component (Aronson, 1999) designed to encourage students to internalize the key message, as described below (see also Table S29). The materials did not treat students as in need of help but instead asked students for their assistance in helping future sixth-grade students.

The first session took advantage of upcoming state tests to address belonging concerns in the context of test-taking and how these can lessen with the help of teachers. One story read:

> "I didn't like taking tests at the beginning of sixth grade . . . I thought I wasn't prepared, and that my teachers and other people would think I wasn't smart. Sometimes when I had to take a test my stomach hurt. But the teachers were really nice. They helped me get better even if I didn't do well at first. Now I know I can trust people here. Teachers are on your side at [school name], and you make friends who help you out."

Students were then asked to write one or two reasons (a) why sixth graders might worry at first about taking tests, (b) why this worry can decline over time, and (c) why students can succeed on tests even if they worry about them. One Black boy wrote, respectively, "May'be they thing they weren't prepared. They thought there teachers might think they're stupid" (Prompt 1), "They tried to do thier best. They got treated with respecct from there teachers" (Prompt 2), and "They tried there best. The teachers are there to help them" (Prompt 3). For more student responses, see Table S29. As noted, no students were excluded based on the presence or quality of their responses to the intervention prompts.

The second intervention session addressed belonging more broadly. One story read:

> "Middle school is scary at first but it gets better. [School name] is big. You have to be more independent and change classes. I worried I wouldn't find my classes, and that I'd forget my locker combination. But [school name] teachers and staff care about you. Once I got lost but the people I asked showed me the way. Even when I got in trouble or didn't do well in class the teachers showed me respect. They are easy to talk to and they listen to what you have to say. I have good friends now at [school name], I get along well with my teachers, and I feel at home here."

Students were then asked to write about why sixth graders might (a) worry at first about whether they belong in middle school and (b) why this worry lessens with time. One Black boy wrote, respectively, "They might thinck they're dumb and everyone will laugh at them. They might get lost and forget their locker combination" (Prompt 1) and "Because their teachers respect you. Because they now have really close friends" (Prompt 2).

The control exercise was identical in form and interactivity but addressed unrelated aspects of the middle school transition—how

students become more interested in state politics (Session 1) and become accustomed to noise and food in the school lunchroom (Session 2) with time.

Dependent measures. The primary outcome was discipline citations obtained from official school records. These were analyzed as an annual average (i.e., using a single aggregate outcome, and corresponding single-level model) and longitudinally across seven grades (citations per grade, from Grades 6 to 12) and six semesters of middle school (citations per semester). Social-psychological climate surveys assessed a standard battery of measures used in other randomized trials of social-psychological interventions in middle school (Cohen et al., 2006, 2009; Cook et al., 2012; Yeager et al., 2014). Three postintervention measures were most relevant to the social-belonging intervention: belonging uncertainty (two items; Walton & Cohen, 2007), worry about stereotype threat (six items; Cohen & Garcia, 2005), and level of social belonging (10 items; Cook et al., 2012; Walton & Cohen, 2007). These are described below. For the full scales, see the online supplemental materials. Belonging uncertainty was the construct the intervention was designed to affect most directly. We expected effects on the other two measures to emerge over time.

Annual average of discipline citations in middle and high school. Discipline citations consisted of the number of officially recorded citations during each school year for which students received a consequence, as determined by classroom teachers or other school staff. For middle school, the primary source of data was detailed paper records for individual students that included the date and description of each citation. These listed approximately 90% of citations for students in each grade. The school district also provided electronic end-of-year counts that included a few additional citations not included in the paper records. Citations in the electronic records did not have descriptions, so they could not be included in analyses by citation type (subjective or objective); however, they were included in analyses of average annual citations. Although the specific date for each subjective and objective citation was available, we aggregated citations by semester (Fall semester: Quarters 1 and 2; Spring semester: Quarters 3 and 4) for the purpose of analysis.

The district-provided discipline records were used to calculate three high-level outcomes: the average number of discipline citations each student received per year over the full 7-year assessment period (from Grades 6–12) and average annual citations received in middle school (the 3-year annual average from Grades 6–8) and high school (the 4-year annual average from Grades 9–12) separately. As in Experiment 1, we used the annual average as the outcome (not the sum total) because a few students left or reentered the school system each year (see Table S32). Using the sum total citations does not change the results.

Subjective and objective citations in middle school. As in Experiment 1, middle-school (Grade 6 to 8) records briefly described each citation. These descriptions allowed us to classify nearly all middle-school citations in the paper records (97%) as either "subjective" or "objective," using the same coding scheme as in Experiment 1 (see Tables S10–S11; citations that did not clearly fit into either category or were labeled "other" could not be classified, and were thus included only in analyses of overall citations). A total of 53% of citations issued by the school to participants over their 3 years of middle school were subjective and 44% were objective (the remaining 3% were unspecified).

High-school records only indicated total citations and did not permit this distinction.

Social-psychological outcomes. All psychological outcomes were assessed once per term, at the start of fall semesters and at the end of spring semesters, for a total of five postintervention assessments (below we report the median Cronbach's alpha for the five assessments for each scale).

Belonging uncertainty. Belonging uncertainty was measured with a two-item, negatively valenced scale (e.g., "Sometimes I feel like I belong at [school name], and sometimes I feel like I don't belong at [school name]"; "When something bad happens, I feel like maybe I don't belong at [school name]"; 1 = *very much disagree,* 7 = *very much agree*; $\propto_{median}$ = 0.55; Walton & Cohen, 2007, 2011). Higher scores indicated greater belonging uncertainty. Because of the low reliability of these two items in this sample (which have previously been used with college students), we also analyzed each item separately as an outcome (see online supplemental materials). Though results were similar for the two single-item analyses, they were stronger when the second, more direct item was used (see online supplemental materials).

Stereotype threat. Stereotype threat was assessed with a six-item, negatively valenced scale (Cohen & Garcia, 2005). There were five 6-point items (e.g., "In school, I worry that people will judge my racial group, based on the behavior or performance of other people in my race," 1 = *very much disagree,* 6 = *very much agree*) and a single, 7-point item ("Sometimes, in school, other people draw negative conclusions about my race's intellectual ability," 1 = *very much disagree*, 7 = *very much agree*). Higher scores indicated greater threat. The scale had good reliability, $\propto_{median}$ = 0.89.

Social belonging. Social belonging was assessed with a 10-item scale (e.g., "I feel like I belong in my school"; 1 = *strongly disagree,* 6 = *strongly agree;* $\propto_{median}$ = 0.82; Cook et al., 2012; Walton & Cohen, 2007). Three items (e.g., "I feel like an outsider at [school name]") were reverse scored and then averaged with the seven positively valenced items. Higher scores indicated higher levels of social belonging. Students' level of belonging and belonging uncertainty have been conceptualized as separate constructs (Walton & Cohen, 2007); consistent with this, the two measures correlated negatively but only modestly, $r_{Pearson's}$ = −0.33, $p < .001$.

Scale calculation. To facilitate comparisons across outcomes, and to ensure items within a scale had the same metric, all item-level scores for each social-psychological outcome at each timepoint were first converted from their original metric to a 0–10 scale before they were averaged with other items at the same timepoint to form a scale average at that timepoint (see Gehlbach, 2006). This was done by subtracting 1 from the original value (e.g., 1 to 7), dividing the result by one less than the total number of response anchors for the item (5, 6, or 7), and then multiplying the second result by 10. This score preserves the original distance between response anchors while ensuring that all items within a scale have the same minimum and maximum values. The final composite score for each outcome can be interpreted as a percentage of the scale maximum (e.g., a score of 7.8 is equivalent to 78% of the scale maximum).

Baseline measures. We had three kinds of baseline covariates: disciplinary behavior, academic performance, and social-psychological measures. To retain participants with missing data on baseline measures, missing values were replaced with the

relevant mean for a participant's Race × Gender group. Where variables were skewed, they were transformed. For positively skewed variables, we tried a square root, natural log, and reciprocal transformation. For negatively skewed variables, we tried a quadratic and cubic transformation. We selected the transformation that resulted in the greatest reduction in skewness for each variable. The online supplemental materials (Tables S35–S36) displays results for the high-level discipline analyses using each possible transformation for the two baseline discipline covariates; treatment effects for Black boys are robust regardless of transformation used.

Baseline disciplinary behavior. We had two baseline measures of disciplinary behavior in the school year immediately prior to the intervention: (a) *fifth-grade discipline citations*, $M = 0.40$, $SD = 1.94$, $Min = 0$, $Max = 19$, the total number of citations each student received during fifth grade, the school year prior to intervention, as in Experiment 1 (although we also had citation data from third and fourth grades, these were uniformly low and so not used in analyses); and (b) *fifth-grade teacher ratings of classroom behavior, $M = 3.14$, $SD = 0.70$, $Min = 1.29$, $Max = 4$*, reflecting fifth-grade teachers' evaluations of students' classroom behavior as "excellent," "good," "satisfactory," or "poor" along 17 items (e.g., "follows classroom rules," "respects authority," "respects peers," "displays self-control"; see online supplemental materials for the full scale). For teacher ratings, we assigned a numerical (1 = *poor* to 4 = *excellent*) rating to each text response such that higher values reflected better classroom behavior. We then averaged the 17 items. The scale had good reliability ($\propto = 0.97$) and was modestly negatively correlated with total fifth grade-citations, $r_{Spearman} = -.39$, $p < .001$.

Fifth-grade discipline citations were positively skewed, skewness = 7.82, $p < .001$. A log transformation (natural log of the variable + 1) reduced but did not eliminate the skew, skewness = 3.94, $p < .001$. The resulting variable had a mean of 0.14, $SD = 0.44$, $Min = 0$, $Max = 3.00$. This was the same transformation applied to baseline discipline citations in Experiment 1.

Fifth-grade teacher ratings were negatively skewed, skewness = −0.60, $p = .005$. A cubic transformation eliminated the skew, skewness = 0.09, $p = .66$. The resulting variable had a mean of 35.39, $SD = 19.34$, $Min = 2.17$, $Max = 64$.

Baseline social-psychological measures. We did not have fifth-grade measures of social belonging and belonging uncertainty; indeed, in referring to perceptions of elementary rather than middle school, these would have been less relevant. However, the climate assessment administered at the start of the sixth grade (prior to the first intervention session) included three relevant psychological constructs (see analytic approach described below), among other scales.

The first was self-reported *stereotype threat* (six items, five 6-point items and a single 7-point item) identical to the dependent variable scale, $\propto = 0.88$. Although the baseline assessment did not include either belonging measure, it did assess *school trust*, students' perception that they were treated fairly in middle school (five 6-point items; e.g., "I am treated fairly by teachers and other adults at my middle school"; see Yeager et al., 2014, $\propto = 0.68$), and *racial identification*, the importance students placed on their racial-ethnic identity, which is predictive of negatively stereotyped students' school experience (Cohen & Garcia, 2005). Racial identification was assessed with two items ("How important is your

racial/ethnic background to you?"; *1 = not at all important, 5 = really important,* and "Race is an important part of who I am"; *1 = very much disagree, 6 = very much agree; ∝ = 0.78*). As was done with the dependent social-psychological variables, item-level scores for all three baseline social-psychological variables were converted to a 0–10 point scale before averaging.

Self-reported stereotype threat was positively skewed, skewness = 1.23, *p* < .001, *M* = 2.30, *SD* = 2.35, *Min* = 0, *Max* = 9.5. A natural log transformation (natural log of the variable + 1), eliminated the skew, skewness = 0.09, *p* = .66, *M* = 0.95, *SD* = 0.70, *Min* = 0, *Max* = 2.35. Both self-reported school trust, skewness = −0.91, *p* < .001, *M* = 8.23, *SD* = 1.40, *Min* = 4, *Max* = 10, and racial identification, skewness = −0.58, *p* = .006, *M* = 6.37, *SD* = 3.05, *Min* = 0, *Max* = 10, were negatively skewed. A cubic transformation eliminated the skew for school trust, skewness = −0.04, *p* = .83, *M* = 603.13, *SD* = 258.37, *Min* = 64, *Max* = 1000, and a quadratic transformation eliminated the skew for racial identification, skewness = 0.18, *p* = .37, *M* = 49.87, *SD* = 34.86, *Min* = 0, *Max* = 100.

Baseline academic performance. We had two measures of baseline academic performance: (a) *fifth-grade academic evaluations,* a composite measure of teachers' evaluations of students' fifth-grade classroom performance in core academic areas and (b) *fifth-grade state test scores,* scores on the statewide standardized test administered in fifth grade.

Fifth-grade teachers evaluated students' performance in core academic areas (language arts, math, social studies, and science) as "excellent," "good," "satisfactory," or "poor" along a variety of dimensions. For each subject area, we transformed the teacher's evaluation into a 4-point scale with higher values representing better evaluations (1 = *poor*, 4 = *excellent*). After averaging scores across subject areas, we standardized scores among students with available data to create an index of overall classroom-based academic performance in fifth grade, *M* = 0, *SD* = 1, *Min* = −2.31, *Max* = 1.55. Students took the state standardized achievement test in September of fifth grade. To create a composite test score measure, we standardized the raw scores on the three sections (math, reading, and writing) and then averaged these, *M* = 0, *SD* = 0.90, *Min* = −2.56, *Max* = 1.67.

For simplicity, we formed a single composite variable of baseline academic performance by averaging the two standardized indicators, fifth-grade academic evaluations and fifth-grade test scores, as they were highly correlated, $r_{\text{Pearson's}}$ = .71, *p* < .001. The mean of the composite variable was 0 (*SD* = 0.88, *Min* = −2.36, *Max* = 1.46). Though the measure was negatively skewed, we did not transform it as it was not significantly so (skewness = −0.40, *p* = .055), and transformations appropriate for negative skew made the variable more rather than less skewed.

Prior to analyses, all baseline covariates were mean-centered within Participant Race × Gender group, as in Experiment 1, so that adjusted means would represent the mean for the average student in a given Race × Gender group. In addition, to create regression coefficients of similar magnitude, transformed scores for fifth-grade teacher ratings, school trust, and racial identification were divided by 100, and transformed scores for stereotype threat were multiplied by 10 (such scaling does not affect *p* values in regression analyses).

Baseline equivalence. As in Experiment 1, we assessed the success of random assignment by regressing each baseline measure on condition (a single treatment vs. control dummy code), for the full sample and each Race × Gender subsample. We used logistic regression for dichotomous race and gender variables; linear regression for fifth-grade teacher ratings and academic performance, and sixth-grade preintervention psychological beliefs; and negative binomial regression for fifth-grade discipline citations. In the full sample, there were no differences between the control condition and the treatment condition on any baseline measure, *z/t*s ≤ 0.75, *p*s > .45, except for racial identification. The treatment condition had marginally lower levels of racial identification than the control condition, M_C = 6.87 versus M_T = 5.88, *b* = −0.98, *t*(135) = −1.91, *p* = .058. Examining each of the four Race × Gender subsamples separately revealed that this relationship was significant only for White boys. For Black boys, the focal group of interest, there were no significant differences between the control and treatment for any of the six baseline measures, |*z/t*|s < 1, *p*s > .33, except for levels of stereotype threat at the start of sixth grade. Black boys reported higher levels of stereotype threat at the beginning of sixth grade in the treatment condition than in the control condition, M_C = 1.74 versus M_T = 3.87, b = 2.13, *t*(23) = 2.24, *p* = .035. Given the number of comparisons tested (32), observing one difference by chance alone is unsurprising. It was also the case that this condition difference was nonsignificant once the baseline stereotype threat variable was corrected for skewness. Additionally, the direction of this difference works against the predicted treatment benefits for Black boys, and analyses of social-psychological outcomes control for this baseline measure. See Table S33.

As in Experiment 1, we additionally tested for equivalency of random assignment using an omnibus test that involved regressing our principal high-level outcome, annual average discipline citations during middle and high school, on the six baseline measures simultaneously (the two fifth-grade discipline measures, the academic performance composite, and the three social-psychological measures from the start of sixth grade). We then regressed the predicted values from this regression model on the condition dummy code. When we did this, there was no main effect of condition, for either the full sample or Black boys specifically, |*t*|s < 0.85, *p*s ≥ .42.

Data analyses. Students were included in analyses (*N* = 137) if they saw materials for either of the two social-belonging (or control) exercises administered in two sessions early in sixth grade (87% saw the assigned randomized materials for both intervention sessions). Analyses are thus intent-to-treat. As in Experiment 1, all students were retained in analyses, regardless of whether they saw materials for one or both sessions and regardless of whether they responded to the essay prompts.

Single-level analyses. For summative discipline outcomes, we used single-level multiple negative binomial regression to test the treatment effect, the same approach used in Experiment 1. As noted in Experiment 1, we use a full-sample analysis for single-level models to test treatment interactions with race and gender. We also used the same set of predictors as in Experiment 1, comprising the main effects of race, gender, and condition and their two-way interactions. Additionally, we included the three-way interaction between race, gender, and condition (the Experiment 1 model fit better without this interaction). In full-sample analyses, the treatment effect (in negative binomial analyses, the ratio of treatment-condition to control-condition citations) pertains

to the Race × Gender group defined as the reference category (the group coded 0 for both race and gender dichotomous variables). As in Experiment 1, for each analysis involving single-level discipline outcomes, we report 95% confidence intervals for the incident rate ratios associated with the log-count coefficients of the relevant simple effect, and also express incident rate ratios as a percentage change relative to the reference category (i.e., in most analyses, the control condition for a specific Race × Gender group).

Longitudinal analyses. Unlike Experiment 1, Experiment 2 permitted longitudinal analyses of both discipline citations and psychological beliefs. To assess change over time as a function of condition, multilevel regression models were used to estimate longitudinal trajectories for discipline citations (multilevel negative binomial regression) and for high-level psychological beliefs (multilevel linear regression). As discussed previously, we feature subsample results (a model estimated separately for each Race × Gender group) for these longitudinal analyses to best understand treatment effects for Black boys (parallel full-sample results are presented in the online supplemental materials and are generally consistent). Though subsample adjusted means are depicted in figures for all four Race × Gender groups, the primary focus of the longitudinal results section is on treatment effects for Black boys, the group expected to benefit. This group had the highest number of control-condition citations and was also the only group that exhibited a treatment reduction in the summative single-level analysis. The supplement presents detailed results for the other Race × Gender groups.

We summarize the longitudinal models here and provide technical details (including equations for each model) in the supplement. For discipline trajectories, we first examined changes in annual discipline citations over seven years, from Grades 6 to 12, through both middle and high school. We then examined changes in subjective and objective citations per semester both within and across school years in middle school. For social-psychological trajectories, we examined changes in self-reported belonging uncertainty, stereotype threat, and social belonging over five timepoints in middle school from the end of sixth grade to the end of eighth grade.

For every longitudinal analysis except one, we fit multilevel models that contained two submodels: a Level-1 submodel that predicted changes in the outcome over time for a given student, and Level-2 submodels that predicted key parameters of the Level-1 model as a function of condition, controlling for baseline covariates at Level 2 (Singer & Willett, 2003). The analysis of middle-school discipline trajectories by citation type required a three-level model, with change in the outcome over time specified at Level 1, citation type specified at Level 2 (within participants), and condition and baseline discipline citations specified at Level 3 (between participants). We also included time-varying predictors at Level 1 when these allowed us to test additional hypotheses (e.g., condition effects on within-grade slopes). Though we report and display results for all four Race × Gender groups, Black boys were consistently the only group to show meaningful variation in the outcome and corresponding treatment effects across analyses. As with single-level analyses, we report 95% confidence intervals for relevant simple effects in longitudinal analyses, for incident rate ratios (the exponentiated log-count coefficients from longitudinal discipline analyses) or linear coefficients (the linear coeffi-

cients from longitudinal analyses of social-psychological outcomes).

Calculation of effect sizes. Effect sizes for longitudinal analyses of discipline citations are again expressed as incident rate ratios. For longitudinal analyses of social-psychological outcomes, a standard-deviation equivalent was computed by dividing the residual variance estimate in an unconditional longitudinal model predicting the relevant outcome (one with no independent variables) by the average number of measurements per individual, adding the result to the between-subjects variance estimate in the unconditional model (the random intercept), and taking the square root of the sum. This was more conservative and precise than simply using the standard deviation of the raw outcome. Effect sizes for condition effects on outcome level were computed by dividing the coefficient corresponding to the main effect of condition by this standard deviation equivalent. Effect sizes for condition effects on slopes reflect the absolute change in the relevant social-psychological outcome (slope coefficient representing the change in outcome per unit of Time × Duration of the relevant time period) divided by this standard deviation equivalent.

Inclusion of baseline covariates. To increase the precision of treatment effect estimates, we included baseline measures as covariates in regression models. To obtain the most parsimonious model (and to maintain consistency with Experiment 1), single-level regression analyses of discipline outcomes controlled only for the two fifth-grade measures of discipline, fifth-grade discipline citations and fifth-grade teacher ratings of classroom behavior. Both covariates were mean-centered within the four Race × Gender subgroups to preserve preintervention differences and to avoid inappropriate between-groups comparisons in adjusted means. In longitudinal analyses of discipline outcomes fifth-grade teacher ratings of classroom behavior was not predictive once fifth-grade discipline citations were controlled and, thus, was dropped.

For social-psychological outcomes, we did not have baseline measures of two of the three constructs to use as covariates. However, a wide variety of other variables measured at baseline are conceptually relevant to students' psychological experience in school and, thus, potentially appropriate as covariates that could help reduce error (e.g., Walton & Cohen, 2011). Indeed, each of the baseline measures—the two fifth-grade discipline measures, the composite measure of fifth-grade academic performance, and the three social-psychological measures assessed at the start of sixth grade, stereotype threat, school trust, and racial identification—was predictive in analyses of at least one social-psychological outcome. To increase power for analyses of social-psychological outcomes (which, as self-report scales, were measured less reliably than school discipline reports), while also maintaining a common model to facilitate comparability, two-level longitudinal analyses of social-psychological outcomes retained all six covariates at Level 2; as with the discipline analyses, each of the covariates was mean-centered on the mean for each Race × Gender group.

Results and Discussion

We begin with a high-level view of the results: We first report the effects of social-belonging treatment and its interactions with race and gender on the average number of citations

students received per school year through high school, a total of 7 years. After this initial summative analysis, we use longitudinal subsample analyses (i.e., the other three Race × Gender groups are excluded when fitting the model) to inform how the intervention benefited Black boys, whether by creating an initial benefit that persisted (an intercept effect only) or by altering the rate of change (e.g., by slowing an increase in discipline citations over time or by accelerating a decrease). Next, to inform how the intervention altered recursive dynamics between students and teachers (building on Experiment 1), we focus on more granular patterns within middle school. Specifically, we use additional longitudinal subsample analyses to test whether the social-belonging treatment, as predicted, mitigated rising levels of citations among Black boys for incidents requiring the subjective judgment of teachers in the early years of middle school. We supplement these analyses of discipline citations with longitudinal linear regression analyses of how students' social-psychological experience changed over time in middle school, again using a subsample approach.

High-level discipline outcomes.

Average annual discipline citations in middle and high school in the primary sample. Examining average annual citations in the full-sample model (the model fit using data from all four

Race × Gender groups), the negative-binomial regression yielded significant Race × Condition, $b = 1.60$, $z = 2.41$, $p = .016$, $IRR = 4.97$, and Gender × Condition interactions, $b = 1.31$, $z = 2.36$, $p = .018$, IRR = 3.69, and a marginal Race × Gender × Condition interaction, $b = -1.81$, $z = -1.95$, $p = .051$, IRR = 0.16 (see Figure 3). As expected, these interactions were driven by Black boys, who received 65% fewer citations per year, 95% CI [−85%, −15%], over the 7-year assessment period in the social-belonging condition than in the control condition ($M_{adj-C} = 2.86$ vs. $M_{adj-T} = 1.02$), $b = -1.04$, $z = -2.32$, $p = .02$, IRR = 0.35, 95% CI [0.15, 0.85]. Though citations were higher in the treatment than control condition for the other three Race × Gender groups, none of these differences approached significance, in part because they were from a lower base, zs < 1.20, ps > .24. The treatment effect for Black boys was stronger yet in a subsample analysis excluding the other Race × Gender groups ($M_{adj-C} = 2.97$ vs. $M_{adj-T} = 1.06$), $b = -1.03$, $z = -3.02$, $p = .003$, IRR = 0.36, 95% CI [0.18, 0.70].

Reduction of the race-based discipline disparity with White boys. To estimate the reduction in racial disparities in discipline citations, we compared Black boys in each condition with White boys in the control condition (we used the control condition for White boys to ensure any race-based reductions would not be
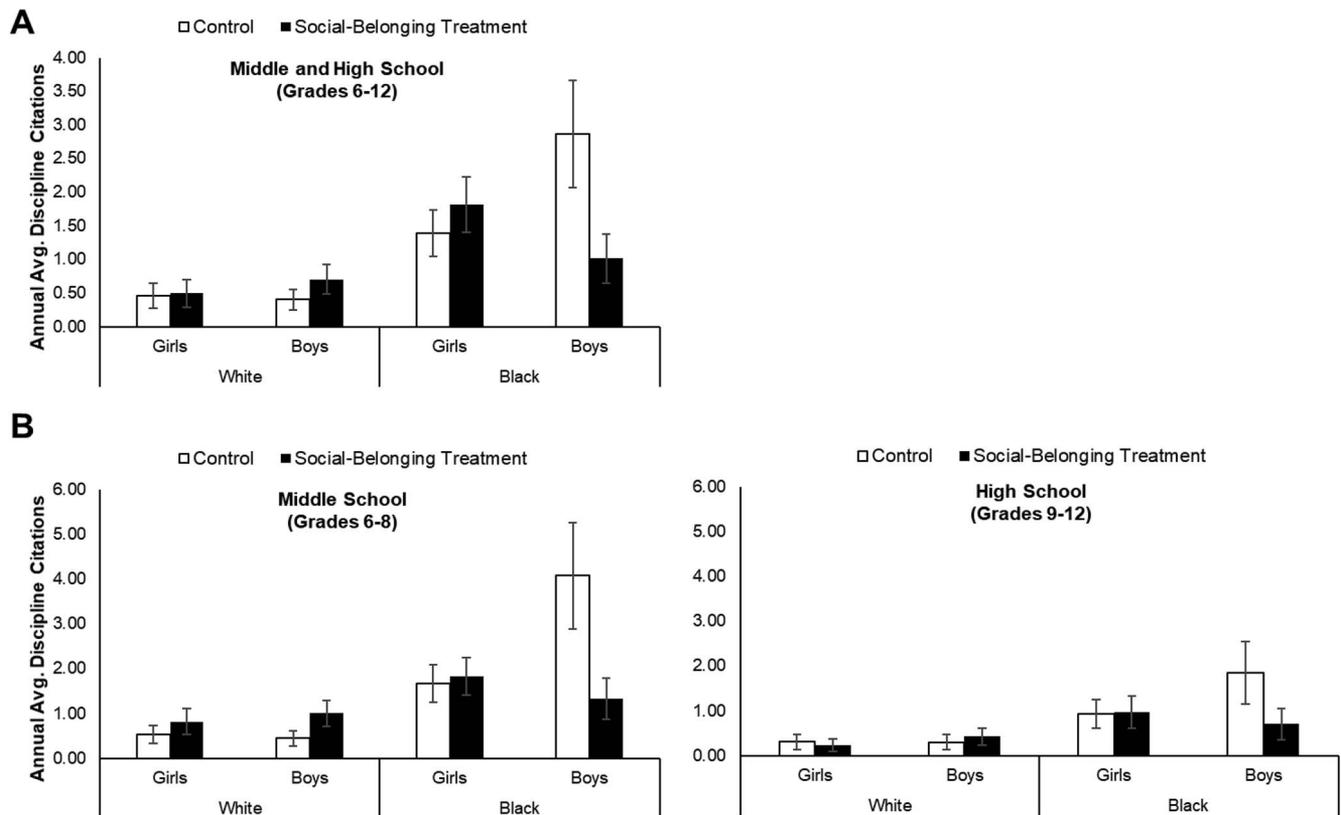

Figure 3. The adjusted average number of annual discipline citations (A) from sixth grade to 12th grade and (B) separately during middle (Grades 6 to 8) and high school (Grades 9 to 12) in Experiment 2, as a function of race, gender, and condition ($N = 137$). Means were obtained from single-level, full-sample negative binomial analyses during each time period and adjusted for preintervention (fifth grade) discipline measures. Error bars represent ±1 SE. For the regression model from which these means were derived, see Table S34.

confounded by the nonsignificant increase among White boys in the treatment condition). Comparing control-condition Black boys with control-condition White boys in a subsample analysis, there was a very large disparity in average annual discipline citations (Ms_{adj} = 3.19 vs. 0.44), b = 1.99, z = 5.30, p < .001, IRR = 7.32, 95% CI [3.50, 15.28], a 632% difference, consistent with national data (Skiba et al., 2011; U.S. Department of Education, 2012; U.S. Government Accountability Office, 2018). The intervention reduced this disparity with control-condition White boys by 75% to a 156% difference in average annual citations, (Ms_{adj} = 1.11 vs. 0.44), b = 0.94, z = 2.22, p = .026, IRR = 2.56, 95% CI [1.12, 5.85].

Average annual discipline citations in middle and high school in the larger sample. The slightly larger sample combining Latino with Black students and Asian with White students (N = 149) produced a similar significant treatment reduction (61%) in average annual citations for negatively stereotyped boys (M_{adj-C} = 2.77 vs. M_{adj-T} = 1.09), b = −0.93, z = −2.33, p = .02, *IRR* = 0.39, 95% CI [0.18, 0.86], and no other group (zs < 1.25, ps ≥ .22; see Table S37). As with the primary sample, the subsample result was stronger than the full-sample result for negatively stereotyped boys, (M_{adj-C} = 2.88 vs. M_{adj-T} = 1.16), b = −0.91, z = −2.82, p = .005, IRR = 0.40, 95% CI [0.21, 0.76]. In addition, a subsample analysis indicated that the race-based discipline disparity relative to control-condition nonstereotyped boys was reduced by 71% from 590% for control-condition negatively stereotyped boys (Ms_{adj} = 3.04 vs. 0.44) to 172% for treatment-condition negatively stereotyped boys (Ms_{adj} = 1.20 vs. 0.44; see Table S41).

Similar treatment effects for negatively stereotyped boys were also obtained when average annual citations in middle and high school were analyzed separately, for both the primary and slightly larger sample (see Tables S38 and S39, respectively; again, no such treatment effects were observed for the other Race × Gender groups in either sample).

Longitudinal discipline outcomes: Middle and high school.

Early discipline citations (a) predict later discipline citations in the control condition but (b) are reduced by the treatment. How could an exercise delivered in two class sessions in the first quarter of sixth grade reduce discipline citations over 7 years? For this to occur, the intervention itself need not be powerful on its own, but needs to be situated within a social system that has the capacity to propagate early change forward in time, as many school contexts often do (Cohen et al., 2009; Goyer et al., 2017). As a first look at this, and taking advantage of the 7-year assessment period, we examined the correlation between discipline citations received in sixth grade and in high school (the 4-year annual average) among control-condition Black boys. This association was exceptionally high, $r_{Spearman's}$ = .81, p = .004 (see Table S42). We used a Spearman's correlation, which assesses the monotonic relationship between the variables through a rank-ordering procedure, because count variables violate the assumptions required for Pearson correlations. In a system like this, early improvement has the potential to cause lasting benefits. (We also examined this correlation in the treatment condition. It was flat, $r_{Spearman's}$ = −.05, p = .87. While this finding is consistent with the interpretation that the treatment interrupted a recursive cycle that led early discipline citations to persist, this estimate should be interpreted with

caution as variance among citations within this condition was low.)

Annual change in discipline citations from sixth to 12th grade. Next we examined the change in annual discipline citations through high school in multilevel negative binomial subsample regression analyses (see Figure 4 and Table S43). These revealed that, for Black boys, the intervention achieved long-term reductions through an initial 78% reduction in sixth-grade citations (M_{adj-C} = 3.18 vs. M_{adj-T} = 0.70), b = −1.51, z = −2.89, p = .004, IRR = 0.22, 95% CI [0.08, 0.62], that then persisted through time, rather than through a deceleration in the annual growth rate of discipline citations (i.e., a reduction in the annual slope). For both conditions, annual citations rose (nonsignificantly) through eighth grade, by 26% per year in the control condition and 49% per year in the treatment condition, IRRs > 1, ps ≥ 0.15; the Middle School Slope × Condition interaction was nonsignificant, b = 0.17, z = 0.46, p = .65, IRR = 1.18. Consequently, treatment-condition Black boys continued to have fewer annual citations (69% reduction) than control-condition Black boys in eighth grade, (M_{adj-C} = 5.07 vs. M_{adj-T} = 1.57), b = −1.18, z = −2.17, p = .03, IRR = 0.31, 95% CI [0.11, 0.89].

Annual citations decreased for both conditions from the end of middle school to the start of high school in ninth grade, by −68% in the control condition and −74% in the treatment condition. Each transition slope (i.e., the change in citations from eighth to ninth grade) was significant, $|z|s$ ≥ 2.78, ps < .006, and there was no Slope × Condition interaction, b = −0.22, z = −0.35, p = .73, IRR = 0.81. Even as citations declined for both conditions with the start of high school, the condition difference first established in sixth grade was maintained in ninth grade and was of a similar magnitude; it was a 72% reduction (M_{adj-C} = 1.69 vs. M_{adj-T} = 0.48), b = −1.26, z = −1.98, p = .048, IRR = 0.28, 95% CI [0.08, 0.99]. Citations were then stable through 12th grade; high school slopes approached zero in both conditions, $IRRs$≈1, ps > 0.75. The treatment reduction in 12th grade was 70% (M_{adj-C} = 1.91 vs. M_{adj-T} = 0.57), b = −1.21, z = −1.87, p = .061, IRR = 0.30, 95% CI [0.08, 1.06] (the marginal significance is likely a consequence of attrition and lower power). In sum, the intervention brought about an initial change in Black boys' outcomes that persisted but did not increase over time.

We then fit the same longitudinal model in the three other Race × Gender subsamples. For these groups (Black girls, White boys, and White girls), there were no significant differences in annual citations between the treatment and control groups at the start or end of either middle school or high school, zs ≤ 1.60, ps ≥ .11, with one exception. There was a significant decrease in citations in 12th grade for White girls, b = −2.15, z = −2.15, p = .032, IRR = 0.12, but this should be interpreted with caution because control-condition citations were already near zero for this Race × Gender group. See the online supplemental materials (and Table S45) for more detailed descriptions of these results.

Discipline outcomes: By citation type in middle school.

Average annual subjective and objective citations in middle school. Next, we examined discipline patterns by citation type within middle school. As noted, this finer level of granularity provides a way to examine the theorized recursive processes that drive disciplinary interactions between students and teachers. Focusing first on high-level patterns by citation type, the treatment produced a similar and significant reduction in both types of
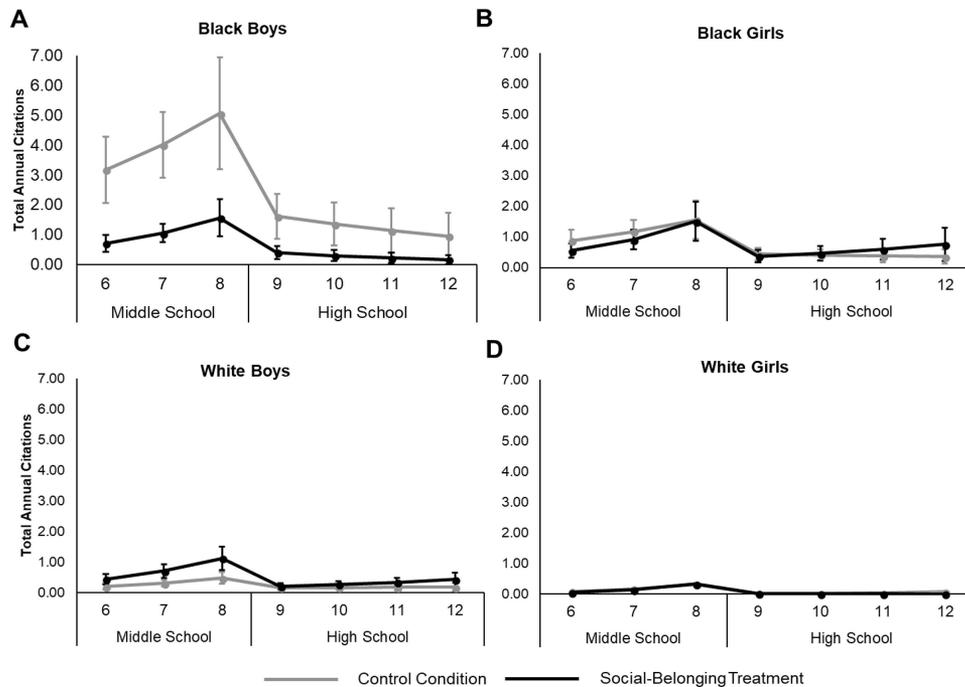
Figure 4. Adjusted longitudinal trajectories in annual discipline citations in Experiment 2 by race, gender, and condition from sixth grade to 12th grade. Each trajectory was derived from a Race × Gender subsample analysis using the multilevel negative binomial regression model presented in Table S43. Means are adjusted for fifth-grade discipline citations, mean-centered within each Race × Gender group. Error bars represent ±1 SE.

citations for Black boys in middle school, when these outcomes were defined as annual averages (as noted previously, citation type was not available for high school). In a subsample analysis, the treatment caused a 77% reduction for average annual subjective citations (M_{adj-C} = 2.79 vs. M_{adj-T} = 0.63), b = −1.49, z = −4.06, p < .001, IRR = 0.23, 95% CI [0.11, 0.46], and a 66% reduction for average annual objective citations (M_{adj-C} = 1.41 vs. M_{adj-T} = 0.49), b = −1.07, z = −2.26, p = .024, IRR = 0.34, 95% CI [0.14, 0.87]. In similar subsample analyses for the other three Race × Gender groups, there were no significant condition differences for either citation type in middle school, $|z|s \leq 1.3$, ps ≥ .19.

Change in subjective and objective citations per semester in middle school. We then examined semester-by-semester change in subjective and objective citations over middle school, within and across grades using multilevel negative binomial regression (a three-level model, see Table S46), focusing on the subsample of Black boys. Despite the similar aggregate (i.e., average annual) effects, these analyses revealed distinct longitudinal patterns for the two citation types.

Within-grade change in subjective citations in sixth and seventh grades. Consistent with the hypothesized deterioration of teacher–student relationships, subjective citations rose within year in the same cyclical pattern for Black boys in the control condition in both sixth and seventh grades (see Figure 5, which also displays subsample longitudinal trajectories for the other three Race × Gender groups). Each year, subjective citations in the control condition began relatively low in the fall semester as students began relationships with new classroom teachers but rose over the

school year, rising by 341% in sixth grade, from 0.43 to 1.90 citations, b = 1.48, z = 2.79, p = .005, IRR = 4.41, 95% CI [1.55, 12.50], and by 125% in seventh grade, from 0.81 to 1.82 citations, b = 0.81, z = 2.56, p = .01, IRR = 2.25, 95% CI [1.21, 4.18]. The social-belonging intervention interrupted this pattern: In the treatment condition the within-year slope did not differ from zero in either grade, zs < 1, ps > .32. Thus, there was a strong treatment reduction in subjective citations in the spring semester each year for Black boys, an 86% reduction in sixth-grade spring (M_C = 1.90 vs. M_T = 0.26), b = −2.00, z = −2.79, p = .005, IRR = 0.14, 95% CI [0.03, 0.55], and an 81% reduction in seventh-grade spring (M_C = 1.82 vs. M_T = 0.34), b = −1.68, z = −3.43, p = .001, IRR = 0.19, 95% CI [0.07, 0.49].

Within-grade change in objective citations in sixth and seventh grades. The within-year rise in subjective citations in sixth and seventh grades experienced by control-condition Black boys was not evident for their objective citations. Though objective citations also rose within school year for the control condition, the increase was not significant in either grade. It was only a 57% increase within sixth grade, b = 0.45, z = 0.65, p = .52, IRR = 1.57, 95% CI [0.40, 6.12], and a 30% increase within seventh grade, b = 0.26, z = 0.66, p = .51, *IRR* = 1.30, 95% CI [0.60, 2.82]. The treatment condition showed similar nonsignificant slopes (see online supplemental materials). The specificity of the effect on subjective and not objective citations suggests that control-condition Black boys did not experience a general decline in self-control over the school year or respond to other seasonal effects. Instead, the pattern points to worsening relationships between Black boys and their teachers as they interacted over the
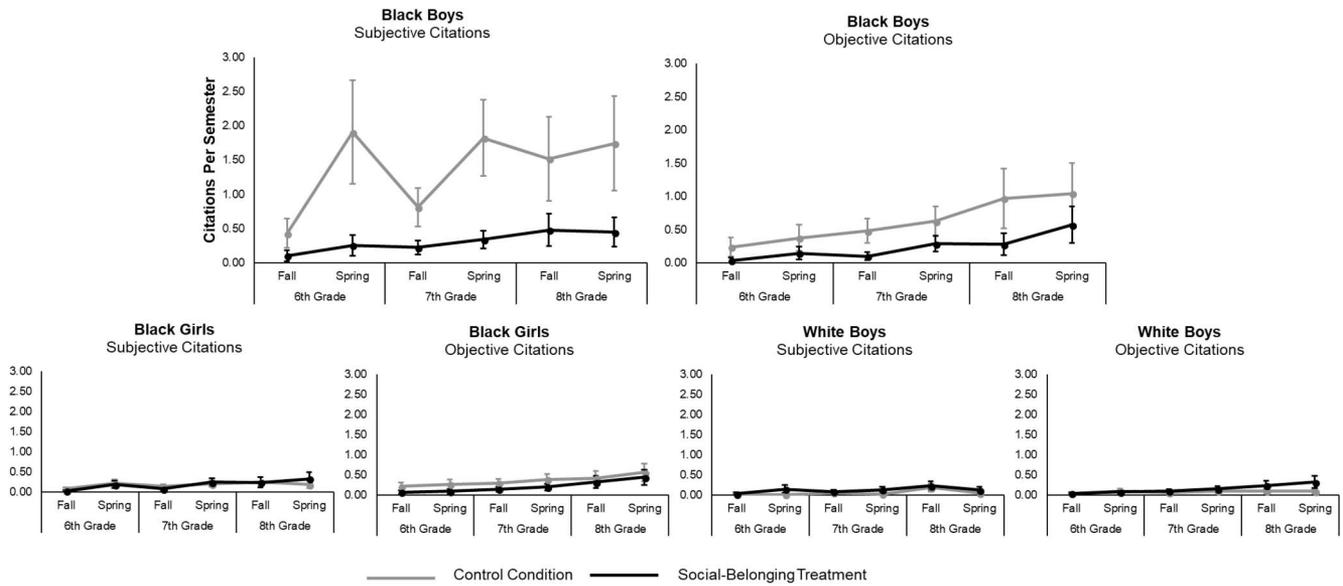
Figure 5. Adjusted discipline citations per semester in middle school (Level 1) in Experiment 2 as a function of citation type (Level 2) and condition (Level 3) for each Race × Gender subsample, modeled longitudinally using multilevel negative binomial regression and adjusted for fifth-grade discipline citations at Level 3. Subjective discipline citations involved greater judgment of the teacher (e.g., "insubordination"). Objective discipline citations involved the application of an objective standard (e.g., "failure to serve detention"). For White girls, the subsample model could not be fit as citations were near zero for both conditions at all timepoints. For the regression model from which these means are derived, see Table S46. Treatment = social-belonging intervention.

school year in sixth and seventh grades. The elimination of this pattern by the treatment points to the contribution of worries about belonging to this process in a context shaped by negative stereotypes.

Across-grade change in citations over middle school. Although only subjective citations rose within grade, both objective and subjective citations rose across middle-school grades (and grade-level teacher teams) in the control condition. This annual increase (between fall semesters) was 101% for objective citations, $b = 0.70$, $z = 1.93$, $p = .054$, IRR = 2.01, 95% CI [0.99, 4.10], and 88% for subjective citations, $b = 0.63$, $z = 2.17$, $p = .03$, IRR = 1.88, 95% CI [1.06, 3.31]. In the treatment condition, these same across-grade slopes were nonsignificant, $zs < 1.65$, $ps \geq 0.11$, even as the percentage change in this condition was from a lower base (due to a treatment reduction in fall semesters; see online supplemental materials).

Lack of reset in the transition to eighth grade. As a result, by eighth grade control-condition Black boys no longer showed a "reset" in subjective citations with the start of a new school year and new classroom teachers. Unlike the transition to seventh grade, when subjective citations (in seventh-grade fall) significantly declined in the control condition by 57% from the spring semester of sixth grade (from 1.90 to 0.81 citations per semester), $b = -0.85$, $z = -2.11$, $p = .035$, IRR = 0.43, 95% CI [0.19, 0.94], subjective citations in the fall semester of eighth grade were, on average, nearly as high as and not significantly different than citations in the spring semester of seventh grade (only a 17% decrease, from 1.82 citations to 1.52 citations), $b = -0.18$, $z = -0.47$, $p = .64$, IRR = 0.83, 95% CI [0.39, 1.78]. Social-

belonging-condition Black boys continued to receive relatively few subjective (and objective) citations across both transitions. The same percentage changes in subjective citations, a 13% decrease in the transition to seventh grade, from 0.26 to 0.22 citations, and a 42% increase in the transition to eighth grade, from 0.34 to 0.48 citations, were nonsignificant for each school-year transition, $zs < 0.65$, $ps > .50$.

Lack of within-grade change in eighth grade. Once in eighth grade, discipline patterns stabilized. Unlike in sixth and seventh grades, there was no significant within-year change in subjective citations in either condition, $|z|s \leq 0.3$, $ps > .75$ (a 15% increase in subjective citations in the control condition and a 7% decrease in subjective citations in the treatment condition).

The rise in both subjective and objective citations in the control condition across grades and teacher teams, and the corresponding high levels of citations in eighth grade, suggests a pattern of worsening behavior over middle school, and foreshadows continued discipline problems in high school.

Longitudinal psychological outcomes: Social-psychological experience in middle school. Multilevel subsample linear regression analyses (see Table S56) examined changes in social-psychological outcomes over middle school. They showed corresponding treatment benefits for the subsample of Black boys. See Figure 6 (which also displays trajectories obtained from subsample longitudinal analyses of the other three Race × Gender groups).

Change in belonging uncertainty over middle school. A reduction in belonging uncertainty for Black boys of 1.59 scale-points (0.81 standard deviations) in the treatment condition relative to the control condition emerged by the first postintervention
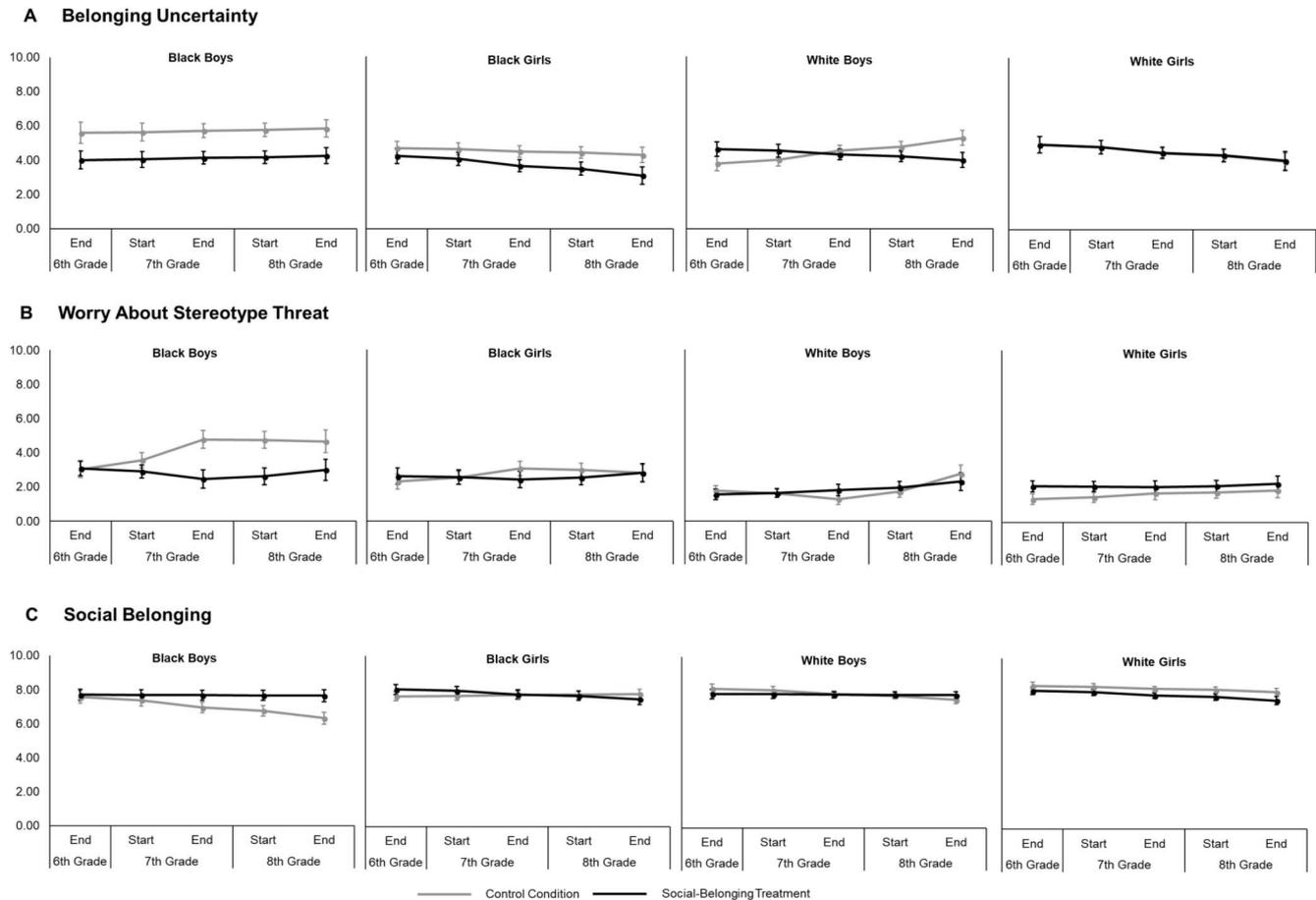
Figure 6. Adjusted longitudinal trajectories over middle school for self-reported (A) belonging uncertainty, (B) worry about stereotype threat, and (C) social belonging. Trajectories reflect adjusted means at five timepoints in middle school for each Race × Gender subsample in Experiment 2 as a function of condition, modeled longitudinally using multilevel linear regression subsample analyses. Means are adjusted for fifth-grade disciplinary measures and academic performance and preintervention (beginning of sixth grade) social-psychological measures at Level 2. For the regression models from which these means are derived, see Table S56. The y-axis represents the full range of each scale after each had been converted to a common metric. Treatment = social-belonging intervention.

assessment, at the end of sixth grade ($M_{\text{adj-C}} = 5.59$ vs. $M_{\text{adj-T}} = 4.00$), $b = -1.59$, $z = -1.93$, $p = .053$, 95% CI $[-3.21, 0.02]$, $d = -0.81$. On average, belonging uncertainty at this time decreased from 56% to 40% of the scale maximum when Black boys received the social-belonging intervention at the beginning of that school year. See Figure 6. The slope for belonging uncertainty over the course of middle school was nonsignificant in both conditions, $zs < 0.42$, $ps \geq .68$, and did not differ by condition, $b = 0.006$, $z = 0.010$, $p = .99$, 95% CI $[-0.96, 0.97]$, $d = 0.01$. Consequently, the reduction in belonging uncertainty established by the end of sixth grade was maintained at the end of seventh grade and in fact was stronger, ($M_{\text{adj-C}} = 5.72$ vs. $M_{\text{adj-T}} = 4.13$), $b = -1.59$, $z = -2.91$, $p = .004$, 95% CI $[-2.66, -0.52]$, $d = -0.81$. It was also present at the end of eighth grade, ($M_{\text{adj-C}} = 5.84$ vs. $M_{\text{adj-T}} = 4.26$), $b = -1.58$, $z = -2.50$, $p = .012$, 95% CI $[-2.82, -0.34]$, $d = -0.80$. Thus, an early reduction in belonging uncertainty persisted through the end of middle school.

For the other three Race × Gender groups, the treatment did not have an initial effect on belonging uncertainty, at the end of sixth grade, $zs \leq 1.30$, $ps \geq .19$. However, treatment effects emerged over time for Black girls and White boys due to effects on slopes. Belonging uncertainty declined marginally over time for treated Black girls, while unexpectedly increased significantly over time for control-condition White boys. By the end of eighth grade, Black girls had marginally lower belonging uncertainty and White boys had significantly lower belonging uncertainty in the treatment condition than in the control condition (see online supplemental materials and Tables S58 to S59).

Change in stereotype threat over middle school. Broader psychological consequences emerged in parallel with rising discipline problems over the course of middle school. Control-condition Black boys' worry about negative stereotypes (e.g., "In school, I worry that people will judge my racial group, based my behavior or performance") rose by 1.74 scale points (absolute increase of 0.81 standard deviations) during the first half of middle

school, from the end of sixth grade to the end of seventh grade, $b = 1.74$, $z = 3.36$, $p = .001$, 95% CI [0.73, 2.76], $d = 0.81$. See Figure 6. Over the same period, treatment-condition scores dropped by 0.62 points (absolute decrease of 0.29 standard deviations), $b = -0.62$, $z = -1.22$, $p = .22$, 95% CI [-1.62, 0.38], $d = -0.29$. Though the treatment-condition decrease did not reach significance, it did result in a significant condition difference in slope in the first half of middle school, $b = -2.36$, $z = -3.35$, $p = .001$, 95% CI [-3.75, -0.98], $d = -1.10$. The treatment reduction in levels of stereotype threat reached significance at the end of seventh grade, when it was 1.08 standard deviations lower in the treatment than control condition ($M_{adj-C} = 4.78$ vs. $M_{adj-T} = 2.47$), $b = -2.31$, $z = -3.08$, $p = .002$, 95% CI [-3.79, -0.84], $d = -1.08$, and persisted subsequently. In the second half of middle school, the slope did not differ from zero in either condition, $zs \leq 0.95$, $ps \geq .34$, as threat levels stabilized; there was no condition difference in slope, $b = 0.65$, $z = 0.84$, $p = .40$. By the end of eighth grade, worry about stereotype threat was still 0.78 standard deviations lower in the treatment condition than in the control condition ($M_{adj-C} = 4.67$ vs. $M_{adj-T} = 3.01$), $b = -1.67$, $z = -1.88$, $p = .06$, 95% CI [-3.40, 0.07], $d = -0.78$.

Change in social belonging over middle school. Concurrently, control-condition Black boys' level of belonging (e.g., "I feel like I belong in my school") declined over the course of middle school by 0.62 scale-points per year (absolute decrease of 0.90 standard deviations over middle school), $b = -0.62$, $z = -3.19$, $p = .001$, 95% CI [-1.00, $-.024$], $d = -0.90$. Levels of social belonging in the treatment condition were stable; the treatment-condition slope over middle school did not differ from zero, $b = -0.03$, $z = -0.18$, $p = .86$, 95% CI [-0.36, 0.30], $d = -0.04$. Consequently, there was a significant condition difference in slope, $b = 0.59$, $z = 2.37$, $p = .018$, 95% CI [0.10, 1.08], $d = 0.85$. Although there was no condition difference in level of social belonging at the end of sixth grade, ($M_{adj-C} = 7.57$ vs. $M_{adj-T} = 7.70$), $b = 0.14$, $z = 0.27$, $p = .79$, 95% CI [-0.86, 1.14], $d = 0.10$, the treatment's prevention of a decline in belonging over time caused levels of belonging to be significantly higher by the end of eighth grade, ($M_{adj-C} = 6.33$ vs. $M_{adj-T} = 7.64$), $b = 1.32$, $z = 2.70$, $p = .007$, 95% CI [0.36, 2.27], $d = 0.95$, equivalent to an increase from 63% to 76% of the scale maximum.

For Black girls, White boys, and White girls there were no significant differences in worry about stereotype threat or in level of social belonging at the first measurement at the end of sixth grade nor at the last measurement at the end of eighth grade, $|z|s \leq 1.70$, $ps \geq 0.09$. See Figure 6 and also Tables S62 to S66.

Summary

The schedule along which the social-psychological measures were assessed was too infrequent to provide nuanced measures of mediation (see Bullock, Green, & Ha, 2010); power is also limited for statistical tests of mediation. However, the results illustrate an important aspect of control-condition Black boys' unfolding experience in middle school, and how this process was altered. When Black boys began sixth grade with an understanding that social challenges and worries about belonging are normal at first in middle school and improve with time, they were able to forestall a within-year rise in subjective citations, first in sixth grade and then again in seventh grade, when interacting with the same

classroom teachers over the course of each school year. They also experienced fewer discipline citations, both subjective and objective, across grades and teacher teams in middle school. By the end of seventh grade, they reported lower worries about being seen and treated in accordance with negative stereotypes. By the end of eighth grade, their level of belonging was higher than controls'. Once a more positive cycle had been established, lower levels of discipline citations continued in eighth grade and then across the transition to high school and through it, 7 years after the intervention. The control-condition pattern in both disciplinary behavior and psychological beliefs was not found for any other Race × Gender group.

The results are consistent with our theory that, here for Black boys, a cycle of student behavior, negative teacher disciplinary judgments, student threat and vigilant reactions, and further teacher responses can become self-perpetuating (Okonofua, Walton et al., 2016). Yet this cycle was interrupted when students began middle school with an adaptive narrative for making sense of early social adversities. The intervention was randomized to students and altered their behavior to initiate improvement, illustrating their agency to shape outcomes in this system. But the intervention's effects did not stay isolated in the minds of students. Instead, the intervention altered patterns of interaction between students and teachers to transform students' school experience.

General Discussion

The present research sheds light on both a social system and cycles of psychological processes that occur among students and teachers within it. In so doing, it highlights a powerful mechanism of social inequality. Black and Latino boys face school contexts replete with negative stereotypes about both their intellectual ability and their behavior (Skiba et al., 2011; Steele, 2010). We theorized that these stereotypes can contribute to early worries about belonging and to negative interactions with teachers, particularly as students and teachers interact over time (Okonofua, Walton et al., 2016). Indeed, more serious and enduring discipline problems can begin as relatively minor incidents, including behaviors perceived by teachers as defiant or disrespectful. Yet even minor initial encounters may confirm fears in negatively stereotyped students about their treatment and social standing in school. Experiences of defiance or disrespect from students can also undermine teachers' experience in school and precipitate burnout (Friedman, 1995; Hastings & Bham, 2003), and fuel further stereotypic judgments of students (e.g., Kunesh & Noltemeyer, 2015; Okonofua & Eberhardt, 2015). Together, negative early encounters, threat, and stereotypes can give rise to a cycle that magnifies and propagates negative outcomes for both parties forward in time.

As interactions and inferences repeat and worsen, teachers and students become stuck in a negative dynamic that is difficult for either to reverse on their own. The cycle gains momentum from its consequences and can cause inequality in children's experience to persist grade after grade for years. Previous work has documented the long-term effects of early behavior problems in preschool (U.S. Department of Education, 2014) and the transition to kindergarten (Myers & Pianta, 2008). A contribution of Experiment 2 in particular is to document the longitudinal trajectory of this process as it emerges in middle school and extends through high school absent intervention to produce sustained high levels of disciplinary

citations for negatively stereotyped boys. Yet a small change at the beginning of this cycle can grow to produce a large cumulative improvement in students' outcomes.

An important contribution of this research is to complement traditional accounts of social inequality, which often locate the problem in fundamental deficits in resources or opportunities on the one hand or in individuals' character or capacities on the other (Walton & Wilson, 2018). Indeed, a very strong correlation between discipline citations in sixth grade and in high school such as that observed in the control condition in Experiment 2 might commonly be interpreted as evidence of a fixed deficit in students, such as an aggressive disposition or a lack of self-control skills. Yet such an interpretation is not supported by the data. The interventions tested here did not address either structural factors or personal deficits directly. Nor did they directly address teachers' perceptions or behaviors, although these are also important (Okonofua, Paunesku et al., 2016). Instead, the results highlight how experiences of identity threat can give rise to recursive psychological and social-relational cycles that undermine outcomes over time. In Experiment 1, an identity-safety intervention, which combined social-belonging, growth-mindset, and values-affirmation interventions, both reduced discipline citations among negatively stereotyped boys across seventh and eighth grades and reduced the odds that a first subjective citation in seventh grade led to another subjective citation in seventh grade or to any citation in eighth grade. In Experiment 2, the social-belonging intervention alone, which conveyed that social challenges in the transition to middle school are normal and improve with time, reduced negative patterns of interaction between Black boys and teachers in sixth and seventh grades, mitigating a within-year rise in subjective citations each year. It also improved the trajectory of Black boys' psychological experience in middle school, and ultimately lowered disciplinary citations through the end of high school. In both experiments, the interventions addressed worries about belonging, inclusion, and growth that arise from awareness of negative stereotypes in the sociocultural context. Their effect was to alter recursive patterns of interaction between teachers and students. Awareness of societal disadvantage has psychological consequences that, in turn, gives rise to unequal cycles of behavior, interaction, and relationships which, in effect, reproduce social inequality (see also Claro, Paunesku, & Dweck, 2016; Yeager, Walton et al., 2016; Walton & Wilson, 2018).

In randomizing students to condition, the design isolates students' agency to alter this system of interactions. In targeting students' experience of identity threat, the studies impugn identity threat in contributing to high rates of school discipline problems among boys of color. From an applied perspective, the results identify a theory-based class of targeted interventions to mitigate persistent discipline problems. Finally, the results show how a brief psychological exercise can transform individuals' lives over very long periods of time—by altering important patterns of interaction—when this exercise is appropriately targeted, timed, and contextually situated (Cohen, Garcia, & Goyer, 2017).

Yet there is much we do not know, both about the unfolding of discipline patterns absent intervention and the role of targeted identity-safety exercises in mitigating these patterns. Indeed, in finding large reductions in discipline citations over years, a significant contribution of the present research is to point toward directions for new research. Although both experiments found

evidence for the theorized recursive cycles between teachers and negatively stereotyped boys, it is essential to further probe the psychological and social-relational mechanisms that underlie discipline patterns over time. Future studies may use more direct measures of students' developing experience, for instance, in scenario studies, classroom observations, or daily diary reports in field settings (Sherman et al., 2013; Walton & Cohen, 2011; see also Hall, Schmader, & Croft, 2015). These may show the effects of race, gender, and identity-safety interventions on how students make sense of specific experiences in class (e.g., interactions with teachers) and how students' interpretations affect their classroom behavior and developing high-level beliefs (e.g., belonging in school, trust of teachers, apprehension about stereotypes). It is also important to focus on teachers, who were not the focus of the present studies and whose role in sustaining discipline cycles is thus less well understood (see Okonofua & Eberhardt, 2015). For instance, research may examine how changes in individual students' classroom behavior affect how teachers make sense of and treat those students either in scenario studies or, ideally, in classroom observations or daily teacher reports in real time and over time (cf. Voigt et al., 2017). Future studies could also cross identity-safety interventions for students with interventions that help teachers reframe student misbehavior (Gregory et al., 2016; Okonofua, Paunesku et al., 2016) or that alter teacher practice (see Yeager et al., 2014), to understand the independent and synergistic effects of student- and teacher-focused approaches.

Better understanding the underlying psychological and social-relational processes will further inform theory. Such an understanding may also reveal ways to improve both the content of identity-safety interventions and their timing. An implication of our findings is that social-belonging, values-affirmation, and growth-mindset interventions may function, at least in some ways, as different routes to promote identity safety within school (see Garcia & Cohen, 2012; Walton et al., 2012). Both growth mindset and social belonging interventions alone were able to achieve meaningful reductions in discipline citations. Further understanding the similarities and differences between these approaches, and how they may be improved and/or combined effectively, is an important direction for future research. Additionally, in the present studies, the interventions were implemented early in adolescence and early in the school year, when students' beliefs about themselves, school, and relationships with classroom teachers were developing. Yet in one study the intervention was delivered in sixth grade and in the other in seventh grade. It is not known how elastic discipline-related processes are to intervention at different points, including whether there is a point at which it is too late for such exercises to meaningfully alter cycles already established (see Cook et al., 2012).

Another important question involves the types of school contexts and types of students in which and for whom identity-safety interventions are more and less likely to be effective. As we have noted, psychological interventions are not magic bullets that will work for all students or in all settings. Instead, they address specific psychological processes that function as barriers within contexts that otherwise permit growth and thus help people stymied by these barriers navigate more effectively (Walton & Wilson, 2018; Yeager & Walton, 2011). In this sense, psychological interventions catalyze potential already present in people and contexts but that is latent or underutilized (Cohen et al., 2017;

Cohen & Sherman, 2014; Walton & Spencer, 2009). There is much to be learned from research exploring the intersection of psychological interventions and social contexts, both for psychological theory and for theory about social contexts. Indeed, even relatively well-researched approaches can be ineffective in specific contexts for reasons that are not yet fully understood (Hanselman et al., 2017; Protzko & Aronson, 2016).

By using different schools, geographical locations, and populations, the present research shows that targeted identity-safety interventions can reduce discipline citations across at least some range of school settings. However, it is not clear to what kinds of schools the three participating schools generalize; thus, we make no strong claims of generalization. As discussed below, already efforts are underway to replicate and extend social-belonging intervention in other middle school contexts (Borman, Rozek, Pyne, and Hanselman (2018).

At a broad level, our results and theory imply that interventions to remedy identity threat will be most effective in adequately resourced contexts with room for improvement: those where identity threat is high and thus undermines teacher–student relationships yet where sufficient material and human resources are present to facilitate better outcomes. Threat may be greater in settings in which inequalities are larger or where negatively stereotyped students are more underrepresented (Borman, Grigg, et al., 2018; Hanselman, Bruch, Gamoran, & Borman, 2014; Walton et al., 2015). It may be lower, and thus identity-safety interventions may be less relevant, when negatively stereotyped students have same-race teachers, who may be less likely to evoke identity threat (Dee, 2004; see also Carrell, Page, & West, 2010; Wright, 2016), or in majority-minority contexts where students who face negative stereotypes may experience greater belonging (Oyserman, Brickman, Bybee, & Celious, 2006). Identity-safety interventions may also be ineffective or even counterproductive if racial bias in disciplinary action is deeply entrenched so such that teacher–student relationships cannot improve, or if there are barriers to better teacher–student relationships beyond identity threat. They may also be less relevant if factors beyond teacher–student relationships, such as school-level discipline policies, classroom-level disruptions, or a lack of student self-control skills, drive discipline problems. Additionally, although we have focused on Black and Latino boys, the specific group that faces negative behavioral stereotypes will differ in different sociocultural and institutional contexts. In settings in which Black and Latina girls or boys of other racial groups face strong negative behavioral stereotypes and receive disproportionate discipline citations, for instance, they too may benefit from identity-safety interventions. There may also be contexts in which concerns about belonging pose challenges for all students. Indeed, using the materials developed in Experiment 2, Borman, Rozek, and colleagues (2018) implemented the social-belonging intervention with diverse sixth grade students in 11 public middle schools across a district. Tracking outcomes over sixth grade, this team observed benefits for all students, including reductions in discipline referrals, higher grades, reduced absences, and improved school attitudes. Understanding social contexts better, and whether and how psychological processes in them serve as barriers whose remedy can improve outcomes for some or all students, is an essential direction for future research.

From an applied as well as a theoretical perspective, it is important to understand how targeted exercises to promote identity safety in students interact with other approaches to mitigate extreme disciplinary problems, such as programmatic efforts to promote identity safety (e.g., ethnic studies curricula; Dee & Penner, 2017), school climate interventions (Bradshaw et al., 2009) and, as noted, teacher-focused interventions (Okonofua, Paunesku et al., 2016; see also Gregory et al., 2016). In what contexts is one approach sufficient, and how can we tell which one? In what contexts is each approach necessary, and does each contribute unique value? In highlighting the role of teacher–student interactions, our research suggests that alternative approaches, such as legislation or policy to prohibit suspension or expulsions for certain discipline violations most associated with racial inequalities ("willful defiance"; Assembly Bill 420, 2014) will not be fully effective unless steps to improve teacher–student relationships are also taken.

These are all exciting directions for future research. Pursuing them will deepen our understanding of students' experience of identity threat and how it unfolds over time in heterogeneous real-world school contexts to shape students' outcomes. Simultaneously, this work will provide further insight into how we can bolster identity safety to mitigate disproportionate levels of discipline citations and ultimately facilitate a better trajectory through school and life for children.

References

Aronson, E. (1999). The power of self-persuasion. *American Psychologist, 54,* 875–884. http://dx.doi.org/10.1037/h0088188

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, pp. 1–79). Reading, MA: Addison Wesley.

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on Black college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38,* 113–126. http://dx.doi.org/10.1006/jesp.2001.1491

Assembly Bill 420: Pupil Discipline: Suspensions and Expulsions: Willful Defiance, 48900, State of California § 660 (2014).

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., & Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors, 27,* 166–177. http://dx.doi.org/10.1037/a0029508

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21,* 726–735. http://dx.doi.org/10.1037/0893-3200.21.4.726

Balderas, G. G. (2014). *Objective versus subjective discipline referrals in a school district* (Doctor of Education Doctoral Dissertation). University of Oregon, Eugene, OR.

Belmi, P., Barragan, R. C., Neale, M. A., & Cohen, G. L. (2015). Threats to social identity can trigger social deviance. *Personality and Social Psychology Bulletin, 41,* 467–484. http://dx.doi.org/10.1177/0146167215569493

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78,* 246–263. http://dx.doi.org/10.1111/j.1467-8624.2007.00995.x

Borman, G. D., Grigg, J., Rozek, C. S., Hanselman, P., & Dewey, N. A. (2018). Self-affirmation effects are produced by school context, student engagement with the intervention, and time: Lessons from a district-wide implementation. *Psychological Science, 29,* 1773–1784. http://dx.doi.org/10.1177/0956797618784016

Borman, G. D., Rozek, C. S., Pyne, J. R., & Hanselman, P. (2018). Reappraising academic and social adversity improves middle-school students' academic achievement, behavior, and well-being. Manuscript under review.

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. *Prevention Science, 10,* 100–115. http://dx.doi.org/10.1007/s11121-008-0114-9

Broda, M., Yun, J., Schneider, B., Yeager, D. S., Walton, G. M., & Diemer, M. (2018). Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *Journal of Research on Educational Effectiveness, 11,* 317–338. http://dx.doi.org/10.1080/19345747.2018.1429037

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology, 98,* 550–558. http://dx.doi.org/10.1037/a0018933

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics, 125,* 1101–1144. http://dx.doi.org/10.1162/qjec.2010.125.3.1101

Celious, A., & Oyserman, D. (2001). Race from the inside: An emerging heterogeneous race model. *Journal of Social Issues, 57,* 149–165. http://dx.doi.org/10.1111/0022-4537.00206

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences of the United States of America, 113,* 8664–8668. http://dx.doi.org/10.1073/pnas.1608207113

Cohen, G. L., & Garcia, J. (2005). "I am us": Negative stereotypes as collective threats. *Journal of Personality and Social Psychology, 89,* 566–582. http://dx.doi.org/10.1037/0022-3514.89.4.566

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313,* 1307–1310. http://dx.doi.org/10.1126/science.1128317

Cohen, G. L., Garcia, J., & Goyer, J. P. (2017). Turning point: Targeted, tailored, and timely psychological intervention. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (2nd ed., pp. 657–686). New York, NY: Guilford Press.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324,* 400–403. http://dx.doi.org/10.1126/science.1170769

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self-affirmation and social psychological intervention. *Annual Review of Psychology, 65,* 333–371. http://dx.doi.org/10.1146/annurev-psych-010213-115137

Cohen, G. L., Steele, C. M., & Ross, L. D. (1999). The mentor's dilemma: Providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin, 25,* 1302–1318. http://dx.doi.org/10.1177/0146167299258011

Cook, J. E., Purdie-Vaughns, V., Garcia, J., & Cohen, G. L. (2012). Chronic threat and contingent belonging: Protective benefits of values affirmation on identity development. *Journal of Personality and Social Psychology, 102,* 479–496. http://dx.doi.org/10.1037/a0026312

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83,* 1314–1329. http://dx.doi.org/10.1037/0022-3514.83.6.1314

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92,* 1006–1023. http://dx.doi.org/10.1037/0022-3514.92.6.1006

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20–33. http://dx.doi.org/10.1037/0022-3514.44.1.20

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics, 86,* 195–210. http://dx.doi.org/10.1162/003465304323023750

Dee, T. S., & Penner, E. (2017). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal, 54,* 127–166. http://dx.doi.org/10.3102/0002831216677002

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11,* 315–319. http://dx.doi.org/10.1111/1467-9280.00262

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98,* 198–208. http://dx.doi.org/10.1037/0022-0663.98.1.198

Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology, 34,* 590–598. http://dx.doi.org/10.1037/0022-3514.34.4.590

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science, 17,* 383–386. http://dx.doi.org/10.1111/j.1467-9280.2006.01716.x

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Iver, D. M. (1993). Development during adolescence. The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist, 48,* 90–101. http://dx.doi.org/10.1037/0003-066X.48.2.90

Emerson, K. T. U., & Murphy, M. C. (2015). A company I can trust? Organizational lay theories moderate stereotype threat for women. *Personality and Social Psychology Bulletin, 41,* 295–307. http://dx.doi.org/10.1177/0146167214564969

Ferguson, R. F., Phillips, S. F., Rowley, J. F. S., & Friedlander, J. W. (2015). *The influence of teaching: Beyond standardized test scores: Engagement, mindsets, and agency: A study of 16,000 sixth through ninth grade classrooms.* Cambridge, MA: Retrieved from http://www.agi.harvard.edu/projects/TeachingandAgency.pdf

Finkel, E. J., Slotter, E. B., Luchies, L. B., Walton, G. M., & Gross, J. J. (2013). A brief intervention to promote conflict reappraisal preserves marital quality over time. *Psychological Science, 24,* 1595–1601. http://dx.doi.org/10.1177/0956797612474938

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York, NY: McGraw-Hill.

Friedman, I. A. (1995). Student behavior patterns contributing to teacher burnout. *The Journal of Educational Research, 88,* 281–289. http://dx.doi.org/10.1080/00220671.1995.9941312

Garcia, J., & Cohen, G. L. (2012). A social psychological perspective on educational intervention. In E. Shafir (Ed.), *Behavioral foundations of policy* (pp. 329–350). Princeton, NJ: Russell Sage Foundation and Princeton University Press.

Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project.* Retrieved from http://www.metproject.org/downloads/Preliminary_Finding-Policy_Brief.pdf

Gay, G., & Howard, T. C. (2001). Multicultural education for the 21st century. *Teacher Educator, 36,* 1–16. http://dx.doi.org/10.1080/08878730009555246

Gehlbach, H. (2006). How changes in students' goal orientations relate to outcomes in social studies. *The Journal of Educational Research, 99,* 358–370. http://dx.doi.org/10.3200/JOER.99.6.358-370

Gershenson, S., Hart, C. M. D., Lindsay, C. A., & Papageorge, N. W. (2017). *The long-run impacts of same-race teachers.* Retrieved from

https://aefpweb.org/sites/default/files/webform/42/LongRunMismatch_AEFP.pdf

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student-teacher demographic match on teacher expectations. *Economics of Education Review, 52,* 209–224. http://dx.doi.org/10.1016/j.econedurev.2016.03.002

Gibbons, R. D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H. C., Greenhouse, J. B., . . . Watkins, J. T. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data. Application to the NIMH treatment of Depression Collaborative Research Program dataset. *Archives of General Psychiatry, 50,* 739–750. http://dx.doi.org/10.1001/archpsyc.1993.01820210073009

Gilliam, W. S., Maupin, A. N., Reyes, C. R., Accavitti, M., & Shic, F. (2016). *Do early educators' implicit biases regarding sex and race relate to behavior expectations and recommendations of preschool expulsions and suspensions?* Retrieved from http://ziglercenter.yale.edu/publications/Preschool%20Implicit%20Bias%20Policy%20Brief_final_9_26_276766_5379.pdf

Goff, P. A., Jackson, M. C., Di Leone, B. A., Culotta, C. M., & DiTomasso, N. A. (2014). The essence of innocence: Consequences of dehumanizing Black children. *Journal of Personality and Social Psychology, 106,* 526–545. http://dx.doi.org/10.1037/a0035663

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology, 102,* 700–717. http://dx.doi.org/10.1037/a0026659

Gottman, J., Notarius, C., Markman, H., Bank, S., Yoppi, B., & Rubin, M. E. (1976). Behavior exchange theory and marital decision making. *Journal of Personality and Social Psychology, 34,* 14–23. http://dx.doi.org/10.1037/0022-3514.34.1.14

Goyer, J. P., Garcia, J., Purdie-Vaughns, V., Binning, K. R., Cook, J. E., Reeves, S. L., . . . Cohen, G. L. (2017). Self-affirmation facilitates minority middle schoolers' progress along college trajectories. *Proceedings of the National Academy of Sciences of the United States of America, 114,* 7594–7599. http://dx.doi.org/10.1073/pnas.1617923114

Gray, M. S. (2016). Saving the lost boys: Narratives of discipline disproportionality. *Educational Leadership and Administration: Teaching and Program Development, 27,* 53–80.

Gregory, A., Hafen, C. A., Ruzek, E., Mikami, A. Y., Allen, J. P., & Pianta, R. C. (2016). Closing the racial discipline gap in classrooms by changing teacher practice. *School Psychology Review, 45,* 171–191. http://dx.doi.org/10.17105/SPR45-2.171-191

Gregory, A., Skiba, R. J., & Noguera, P. A. (2010). The achievement gap and the discipline gap: Two sides of the same coin? *Educational Researcher, 39,* 59–68. http://dx.doi.org/10.3102/0013189X09357621

Gregory, A., & Weinstein, R. S. (2008). The discipline gap and African Americans: Defiance or cooperation in the high school classroom. *Journal of School Psychology, 46,* 455–475. http://dx.doi.org/10.1016/j.jsp.2007.09.001

Hall, W., Schmader, T., & Croft, E. (2015). Engineering exchanges: Women's daily experience of social identity threat in engineering cue burnout. *Social Psychological & Personality Science, 6,* 528–534. http://dx.doi.org/10.1177/1948550615572637

Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., . . . Scott-Little, C. (2012). A course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal, 49,* 88–123. http://dx.doi.org/10.3102/0002831211434596

Hanselman, P., Bruch, S. K., Gamoran, A., & Borman, G. D. (2014). Threat in context: School moderation of the impact of social identity threat on racial/ethnic achievement gaps. *Sociology of Education, 87,* 106–124. http://dx.doi.org/10.1177/0038040714525970

Hanselman, P., Rozek, C. S., Grigg, J., & Borman, G. D. (2017). New evidence on self-affirmation effects and theorized sources of heterogeneity from large-scale replications. *Journal of Educational Psychology, 109,* 405–424. http://dx.doi.org/10.1037/edu0000141

Hastings, R. P., & Bham, M. S. (2003). The relationship between student behaviour patterns and teacher burnout. *School Psychology International, 24,* 115–127. http://dx.doi.org/10.1177/0143034303024001905

Hirschfield, P. (2009). Another way out: The impact of juvenile arrests on high school dropout. *Sociology of Education, 82,* 368–393. http://dx.doi.org/10.1177/003804070908200404

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14,* 640–643. http://dx.doi.org/10.1046/j.0956-7976.2003.psci_1478.x

Inzlicht, M., & Kang, S. K. (2010). Stereotype threat spillover: How coping with threats to social identity affects aggression, eating, decision making, and attention. *Journal of Personality and Social Psychology, 99,* 467–481. http://dx.doi.org/10.1037/a0018951

Jacoby-Senghor, D. S., Sinclair, S., & Shelton, J. N. (2016). A lesson in bias: The relationship between implicit racial bias and performance in pedagogical contexts. *Journal of Experimental Social Psychology, 63,* 50–55. http://dx.doi.org/10.1016/j.jesp.2015.10.010

Jung, H., Spjeldnes, S., & Yamatani, H. (2010). Recidivism and survival time: Racial disparity among jail ex-inmates. *Social Work Research, 34,* 181–189. http://dx.doi.org/10.1093/swr/34.3.181

Kenthirarajah, D., & Walton, G. M. (2015). How brief social-psychological interventions can cause enduring effects. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–15). Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781118900772.etrds0026

Kunda, Z. (1999). *Social cognition: Making sense of people.* Cambridge, MA: MIT Press.

Kunesh, C. E., & Noltemeyer, A. (2015). Understanding disciplinary disproportionality: Stereotypes shape pre-service teachers' beliefs about Black boys' behavior. *Urban Education, 2015,* 1–28. http://dx.doi.org/10.1177/0042085915623337

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education.* Cambridge, MA: Harvard University Press.

Locksley, A., Hepburn, C., & Ortiz, V. (1982). Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology, 18,* 23–42. http://dx.doi.org/10.1016/0022-1031(82)90079-8

Losen, D., Hodson, C., Keith, I. M. A., Morrison, K., & Belway, S. (2015). *Are we closing the school discipline gap?* Los Angeles, CA: Retrieved from https://escholarship.org/uc/item/2t36g571

Marigold, D. C., Holmes, J. G., & Ross, M. (2007). More than words: Reframing compliments from romantic partners fosters security in low self-esteem individuals. *Journal of Personality and Social Psychology, 92,* 232–248. http://dx.doi.org/10.1037/0022-3514.92.2.232

Marigold, D. C., Holmes, J. G., & Ross, M. (2010). Fostering relationship resilience: An intervention for low self-esteem individuals. *Journal of Experimental Social Psychology, 46,* 624–630. http://dx.doi.org/10.1016/j.jesp.2010.02.011

McKinnon, J. D., & Bennett, C. E. (2005). *We the people: Blacks in the United States* [Census 2000 Special Reports]. Washington, DC: U. S. Census Bureau.

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. *Science, 244,* 933–938. http://dx.doi.org/10.1126/science.2658056

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., . . . Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 2693–2698. http://dx.doi.org/10.1073/pnas.1010076108

Moyé, L. A., & Deswal, A. (2001). Trials within trials: Confirmatory subgroup analyses in controlled clinical experiments. *Controlled Clini-*

cal Trials, 22, 605–619. http://dx.doi.org/10.1016/S0197-2456(01)00180-5

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research.* New York, NY: Oxford University Press.

Myers, S. S., & Pianta, R. C. (2008). Developmental commentary: Individual and contextual influences on student-teacher relationships and children's early problem behaviors. *Journal of Clinical Child and Adolescent Psychology, 37,* 600–608. http://dx.doi.org/10.1080/15374410802148160

Newcomb, T. M., Koenig, K. E., Flacks, R., & Warwick, D. P. (1967). *Persistence and change: Bennington College and its students after twenty-five years.* New York, NY: Wiley.

Okonofua, J. A., & Eberhardt, J. L. (2015). Two strikes: Race and the disciplining of young students. *Psychological Science, 26,* 617–624. http://dx.doi.org/10.1177/0956797615570365

Okonofua, J. A., Paunesku, D., & Walton, G. M. (2016). Brief intervention to encourage empathic discipline cuts suspension rates in half among adolescents. *Proceedings of the National Academy of Sciences of the United States of America, 113,* 5221–5226. http://dx.doi.org/10.1073/pnas.1523698113

Okonofua, J. A., Walton, G. M., & Eberhardt, J. L. (2016). A vicious cycle: A social-psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science, 11,* 381–398. http://dx.doi.org/10.1177/1745691616635592

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* 4711–4718. http://dx.doi.org/10.1126/science.aac4716

Oyserman, D., Brickman, D., Bybee, D., & Celious, A. (2006). Fitting in matters: Markers of in-group belonging and academic outcomes. *Psychological Science, 17,* 854–861. http://dx.doi.org/10.1111/j.1467-9280.2006.01794.x

Oyserman, D., Terry, K., & Bybee, D. (2002). A possible selves intervention to enhance school involvement. *Journal of Adolescence, 25,* 313–326. http://dx.doi.org/10.1006/jado.2002.0474

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science, 26,* 784–793. http://dx.doi.org/10.1177/0956797615571017

Perry, B. L., & Morris, E. W. (2014). Suspending progress: Collateral consequences of exclusionary punishment in public schools. *American Sociological Review, 79,* 1067–1087. http://dx.doi.org/10.1177/0003122414556308

Pigott, R. L., & Cowen, E. L. (2000). Teacher race, child race, racial congruence, and teacher ratings of children's school adjustment. *Journal of School Psychology, 38,* 177–196. http://dx.doi.org/10.1016/S0022-4405(99)00041-2

Pratto, F., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology, 27,* 26–47. http://dx.doi.org/10.1016/0022-1031(91)90009-U

Protzko, J., & Aronson, J. (2016). Context moderates affirmation effects on the ethnic achievement gap. *Social Psychological & Personality Science, 7,* 500–507. http://dx.doi.org/10.1177/1948550616646426

Purdie-Vaughns, V., & Walton, G. M. (2011). Is multiculturalism bad for African-Americans? Redefining inclusion through lens of identity safety. In L. R. Tropp & R. Mallett (Eds.), *Beyond prejudice reduction to positive intergroup relations* (pp. 159–177). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12319-008

Quinn, D. M. (2017). Racial attitudes of pre K-12 and postsecondary educators: Descriptive evidence from nationally representative data. *Educational Researcher, 46,* 397–411. http://dx.doi.org/10.3102/0013189X17727270

Rafa, A. (2018). *Policy snapshot: Suspension and expulsion.* CO: Retrieved from Denver; https://www.ecs.org/wp-content/uploads/Suspension_and_Expulsion.pdf

Rattan, A., Savani, K., Chugh, D., & Dweck, C. S. (2015). Leveraging mindsets to promote academic achievement: Policy recommendations. *Perspectives on Psychological Science, 10,* 721–726. http://dx.doi.org/10.1177/1745691615599383

Rattan, A., Savani, K., Komarraju, M., Morrison, M. M., Boggs, C., & Ambady, N. (2018). Meta-lay theories of scientific potential drive underrepresented students' sense of belonging to science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology, 115,* 54–75. http://dx.doi.org/10.1037/pspi0000130

Raudenbush, S. W., & Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6,* 387–401. http://dx.doi.org/10.1037/1082-989X.6.4.387

Ritter, J. A., & Taylor, L. J. (2011). Racial disparity in unemployment. *The Review of Economics and Statistics, 93,* 30–42. http://dx.doi.org/10.1162/REST_a_00063

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development.* New York, NY: Holt, Rinehart & Winston. http://dx.doi.org/10.1007/BF02322211

Rothwell, P. M. (2005). Treating individuals 2. Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *Lancet, 365,* 176–186. http://dx.doi.org/10.1016/S0140-6736(05)17709-5

Rubinstein, R. S., Jussim, L., & Stevens, S. T. (2018). Reliance on individuating information and stereotypes in implicit and explicit person perception. *Journal of Experimental Social Psychology, 75,* 54–70. http://dx.doi.org/10.1016/j.jesp.2017.11.009

Sadler, M. S., Correll, J., Park, B., & Judd, C. M. (2012). The world is not black and white: Racial bias in the decision to shoot in a multiethnic context. *Journal of Social Issues, 68,* 286–313. http://dx.doi.org/10.1111/j.1540-4560.2012.01749.x

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39,* 590–598. http://dx.doi.org/10.1037/0022-3514.39.4.590

Salvatore, J., & Shelton, J. N. (2007). Cognitive costs of exposure to racial prejudice. *Psychological Science, 18,* 810–815. http://dx.doi.org/10.1111/j.1467-9280.2007.01984.x

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., . . . Cohen, G. L. (2013). Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology, 104,* 591–618. http://dx.doi.org/10.1037/a0031495

Simmons, R. G., Black, A., & Zhou, Y. (1991). African-American versus white children and the transition into junior high school. *American Journal of Education, 99,* 481–520. http://dx.doi.org/10.1086/443995

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195152968.001.0001

Skiba, R. J., Horner, R. H., Chung, C., Rausch, M. K., May, S. L., & Tobin, T. (2011). Race Is not neutral: A national investigation of African American and Latino disproportionality in school discipline. *School Psychology Review, 40,* 85–107.

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review, 34,* 317–342. http://dx.doi.org/10.1023/A:1021320817372

Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., & Chung, C.-G. (2008). Achieving equity in special

education: History, status, and current challenges. *Exceptional Children, 74,* 264–288. http://dx.doi.org/10.1177/001440290807400301

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85,* 571–581. http://dx.doi.org/10.1037/0022-0663.85.4.571

Snyder, M., Tanke, E. D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology, 35,* 656–666. http://dx.doi.org/10.1037/0022-3514.35.9.656

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629. http://dx.doi.org/10.1037/0003-066X.52.6.613

Steele, C. M. (2010). *Whistling Vivaldi: How stereotypes affect us and what we can do.* New York, NY: Norton.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811. http://dx.doi.org/10.1037/0022-3514.69.5.797

Steele, D. M., & Cohn-Vargas, B. (2013). *Identity safe classrooms.* Thousand Oaks, CA: Corwin.

Twenge, J. M., Baumeister, R. F., Tice, D. M., & Stucke, T. S. (2001). If you can't join them, beat them: Effects of social exclusion on aggressive behavior. *Journal of Personality and Social Psychology, 81,* 1058–1069. http://dx.doi.org/10.1037/0022-3514.81.6.1058

Tyler, T. R. (1990). *Why people obey the law: Procedural justice, legitimacy, and compliance.* New Haven, CT: Yale University Press.

U.S. Department of Education, Office for Civil Rights. (2012). *Civil rights data collection.* Retrieved from https://ocrdata.ed.gov/

U.S. Department of Education, Office for Civil Rights. (2014). *Data snapshot: School discipline.* Retrieved from https://ocrdata.ed.gov/downloads/crdc-school-discipline-snapshot.pdf

U.S. Government Accountability Office. (2018). *Discipline disparities for Black students, boys, and students with disabilities.* (Publication No. GAO-18–258). Retrieved from https://www.gao.gov/products/GAO-18-258

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., . . . Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences of the United States of America, 114,* 6521–6526. http://dx.doi.org/10.1073/pnas.1702413114

Walton, G. M., & Brady, S. T. (2017). The many questions of belonging. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation* (2nd ed., pp. 272–293). New York, NY: Guilford Press.

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality & Social Psychology, 92,* 82–96. http://dx.doi.org/10.1037/0022-3514.92.1.82

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331,* 1447–1451. http://dx.doi.org/10.1126/science.1198364

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform

women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology, 107,* 468–485. http://dx.doi.org/10.1037/a0037461

Walton, G. M., Paunesku, D., & Dweck, C. S. (2012). Expandable selves. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 141–154). New York, NY: Guilford Press.

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20,* 1132–1139. http://dx.doi.org/10.1111/j.1467-9280.2009.02417.x

Walton, G. M., & Wilson, T. D. (2018). Wise interventions: Psychological remedies for social and personal problems. *Psychological Review, 125,* 617–655. http://dx.doi.org/10.1037/rev0000115

Warikoo, N., Sinclair, S., Fei, J., & Jacoby-Senghor, D. (2016). Examining racial bias in education: A new approach. *Educational Researcher, 45,* 508–514. http://dx.doi.org/10.3102/0013189X16683408

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology, 10,* 109–120. http://dx.doi.org/10.1016/0022-1031(74)90059-6

Wright, A. C. (2016). *Teachers' perceptions of students' disruptive behavior: The effect of racial congruence and consequences for school suspension.* Bellingham, WA: Western Washington University. Retrieved from https://www.aeaweb.org/conference/2017/preliminary/paper/SdzYZefs

Yeager, D. S., Dahl, R., & Dweck, C. S. (2018). Why interventions to influence adolescent behavior often fail but could succeed. *Perspectives on Psychological Science, 13*(1), 101–122. http://dx.doi.org/10.1177/1745691617722620

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., . . . Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General, 143,* 804–824. http://dx.doi.org/10.1037/a0033906

Yeager, D. S., Purdie-Vaughns, V., Hooper, S. Y., & Cohen, G. L. (2017). Declining institutional trust among racial and ethnic minority adolescents: A consequence of procedural injustice and a cause of life-span outcomes. *Child Development, 88,* 658–676. http://dx.doi.org/10.1111/cdev.12697

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research, 81,* 267–301. http://dx.doi.org/10.3102/0034654311405999

Yeager, D. S., Walton, G. M., Brady, S. T., Akcinar, E. N., Paunesku, D., Keane, L., . . . Dweck, C. S. (2016). Teaching a lay theory before college narrows achievement gaps at scale. *Proceedings of the National Academy of Sciences of the United States of America, 113,* E3341–E3348. http://dx.doi.org/10.1073/pnas.1524360113